IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 99, No. 1, Month 2099, pp. 1~1x
ISSN: 2252-8938, DOI: 10.11591/ijai.v99.i1.pp1-1x

Vision Based Drone-Human Differentiation for Aerial Security System

Ezha Tri Saputra [1], Ahmad Riyad Firdaus [1], Daipansyah Arya Saputra[1], Hendawan Soebhakti[1], Emelia Rosari Siregar[1]
[1]Department of Electronics Engineering, Politeknik Negeri Batam, Batam, Indonesia

Article Info

Article history:

Received month dd, yyyy
Revised month dd, yyyy
Accepted month dd, yyyy

Keywords:

Drone Detection
Human Detection
YOLO11s
Real-Time Surveillance
Aerial Security

ABSTRACT

The increasing use of unmanned aerial vehicles (UAVs) has raised serious security concerns, particularly in distinguishing unauthorized drones from humans in restricted areas. This study proposes a real-time vision-based detection framework using a lightweight YOLO11s model to perform accurate drone and human detection from video streams. A hybrid dataset consisting of 2,082 original images was expanded to 5,440 images through structured data augmentation to enhance model robustness and generalization capability. The trained model was evaluated through offline video testing and real-time deployment scenarios. Experimental results demonstrate that the proposed system achieves a precision of 0.912, recall of 0.903, and an F1-score of 0.907, indicating balanced and reliable detection performance. Furthermore, real-time experiments on heterogeneous hardware platforms show that GPU acceleration significantly improves system efficiency, achieving 12.4 frames per second (FPS) with an inference time of 81.0 ms, compared to 5.2 FPS and 191.8 ms on CPU. These findings confirm the effectiveness and practicality of the proposed framework for real-world aerial surveillance and security applications.

Corresponding Author:

Ezha Tri Saputra
Department of Electronics Engineering, Politeknik Negeri Batam, Batam
Jl. Ahmad Yani, Batam Center, Batam, Kepulauan Riau 29461
Email: Trisaputraezha@gmail.com

1. INTRODUCTION

The increasing accessibility of unmanned aerial vehicles (UAVs) has accelerated their adoption across multiple domains such as surveillance, smart cities, logistics, and emergency response[1], [2]. However, this widespread usage has also raised serious concerns regarding unauthorized drone operations in protected or monitored areas[3], [4]. In practical scenarios, security systems are often required to monitor specific zones and identify potential aerial threats in real time. As a result, the ability to accurately detect drones and differentiate them from other relevant objects, particularly humans within the monitored zone, has become a critical research problem in aerial security systems[5], [6].

In this context, vision-based detection systems utilizing video streams have attracted considerable research interest owing to their compatibility with existing camera infrastructures and their ability to support real-time operation [7]. During dataset preparation, video frames are extracted and manually filtered to ensure data quality and class balance during dataset preparation[8]. Compared to radar and acoustic sensors, camera-based approaches provide richer spatial information for object classification and tracking. Recent advancements in deep learning have further improved detection accuracy for small and fast-moving objects such as drones[9], [10], [11].

Building upon these developments, several studies have explored drone detection using convolutional neural networks (CNNs) and transformer-based architectures[12], [13], [14]. Among various approaches, YOLO-based detectors are widely recognized for their high speed and suitability for real-time applications[15]. Recent research utilizing YOLOV8,YOLOV9,YOLOV10, and YOLO-NAS has demonstrated superior performance in detecting UAVs from video streams[16], [17], [18]. Nevertheless, most existing works focus solely on drone detection and do not explicitly address human differentiation[19], [20]. Consequently, although numerous studies have addressed UAV detection, limited attention has been given to joint drone–human differentiation under real-time video scenarios, particularly using lightweight models suitable for deployment on heterogeneous hardware platforms.

From an implementation perspective, the transition from offline video analysis to real-time deployment is a crucial step for practical security systems, as traditional object detection approaches are commonly evaluated in offline settings before being adapted to streaming perception frameworks that address inference latency, frames per second (FPS), and online processing requirements[21]. In line with this, previous studies have validated detection models using recorded video sequences prior to implementing real-time streaming detection, particularly in UAV and surveillance applications where efficient live video processing is essential[22]. This staged evaluation strategy enables comprehensive performance evaluation, model optimization, and detailed analysis of inference time and frames per second (FPS) prior to field deployment[23], [24]. Accordingly, in this study, offline video testing is first conducted to verify detection accuracy and system stability, followed by real-time deployment to evaluate system responsiveness under live video streams[22], [25].

In addition, high-quality and diverse datasets play a crucial role in improving model generalization[26].To address this, data augmentation techniques, including flipping, rotation, brightness adjustment, and noise injection, are widely applied to enhance model robustness against environmental variations[27].Furthermore, public datasets obtained from platforms such as Roboflow have significantly contributed to object detection research. However, studies integrating self-collected datasets with public datasets specifically for drone–human differentiation remain limited, highlighting the need for more comprehensive dataset fusion strategies[28].Therefore, this framework provides a practical solution for real-time aerial security monitoring systems that require fast, reliable, and hardware-efficient object detection.

To address the identified research gaps, this study proposes a vision-based detection framework using the YOLO11s model to distinguish drones and humans from video streams and deployment real-time monitoring. Specifically, YOLO11s is selected due to its lightweight architecture, fast inference capability, and suitability for deployment on both CPU and GPU platforms while maintaining competitive detection accuracy. In this work, the dataset consists of 2,082 original images, which are expanded to 5,440 images through a customized augmentation pipeline. Subsequently, the trained model is initially evaluated using offline video data and later deployed for real-time detection scenarios. Model performance is assessed using precision, recall, and F1-score metrics.

Finally, the novelty of this study lies in the integration of hybrid datasets, structured augmentation strategies, and a lightweight YOLO11s architecture for real-time drone–human differentiation. Moreover, performance evaluation is conducted on both CPU (Ryzen 7 5000 Series) and GPU (NVIDIA RTX 3060) platforms to analyze inference speed and real-time FPS performance. This dual-platform evaluation provides insights into system scalability and hardware efficiency for real-world deployment. Overall, the proposed approach demonstrates improved detection reliability under diverse environmental conditions, making it suitable for real-time aerial security applications.

2. METHOD

2.1 You Only Look Once (YOLO) Architecture

You Only Look Once (YOLO) is a single-stage object detection framework that reformulates object detection as a regression problem, directly predicting bounding box coordinates and class probabilities from full images in a single forward pass[29]. Unlike traditional two-stage detectors, YOLO performs detection and classification simultaneously, resulting in significantly faster inference speed and enabling real-time object detection applications. YOLO employs a unified neural network architecture that reasons globally about the entire image, reducing background false positives and improving detection robustness across diverse environments[29]. Recent studies highlight that YOLO has undergone continuous architectural evolution, introducing multiple optimized variants to enhance detection accuracy, speed, and computational efficiency for real-world deployment[30]. Recent YOLO variants, including YOLOv8, YOLOv9, YOLOv10, and YOLO-NAS, demonstrate improved detection accuracy and real-time performance, especially for small and fast-moving objects. These gains result from optimized backbone architectures, enhanced feature fusion, and advanced loss functions that reduce inference latency.

Figure 1 This figure illustrates the overall structure of the YOLO object detection framework. The input image, resized to 640 × 640 pixels, is processed through the backbone network consisting of convolutional and pooling layers to extract hierarchical feature representations. These features are then forwarded to the neck module based on Feature Pyramid Network (FPN) and Path Aggregation Network (PANet) for multi-scale feature fusion. Finally, the detection head outputs bounding box coordinates, objectness scores, and class probabilities, enabling simultaneous localization and classification of drones and humans in real-time.

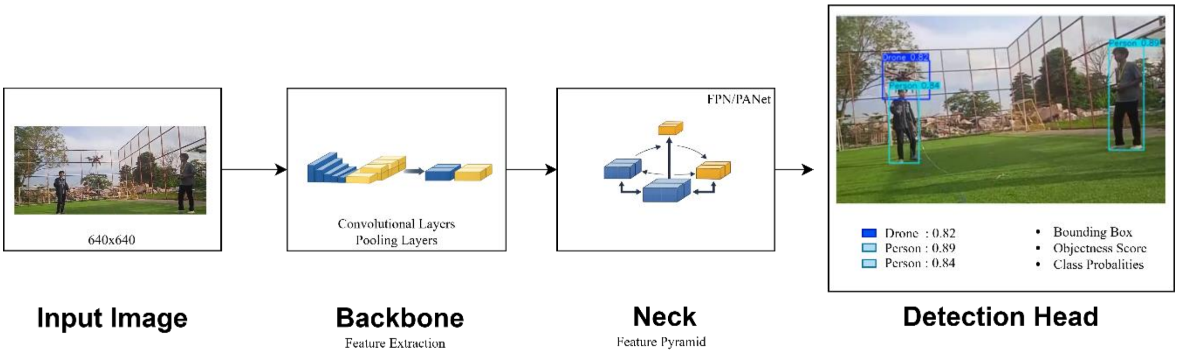


Figure 1. YOLO Architecture

2.2 Experimental Setup and Training configuration

The experimental setup and system configuration used in this study are summarized in Table 1. The model is trained for 100 epochs with an input image resolution of 640 × 640 pixels and a batch size of 16 to ensure stable convergence and efficient memory utilization. The Stochastic Gradient Descent (SGD) optimizer with fixed learning rate and momentum is employed to maintain consistent and reliable optimization behavior throughout the training process.

All training and evaluation procedures are implemented using the Ultralytics YOLO framework with the YOLO11s model. Experiments are conducted on a standardized hardware platform consisting of an AMD Ryzen 7 5000 Series CPU and an NVIDIA GeForce RTX 3060 Laptop GPU to ensure fair performance comparison and experimental reproducibility. Model performance is assessed using widely adopted object detection metrics, including Precision, Recall, mAP@50, and mAP@50–95, to comprehensively evaluate detection accuracy and localization quality.

Table 1. Experimental setup and system configuration for YOLO model training and evaluation.

Category	Component	Specification
Training Parameter	Epoch	100
	Input Image Size	640 × 640 pixels
	Batch Size	16
	Optimizer	Stockhastic Gradient Descent(SGD)
	Learning Rate	Fixed
	Momentum	Fixed
	Framework	Ultralytics YOLO
	Model	YOLO11s
Hardware	Device	GPU(device=0)
	CPU	Ryzen 7 5000 Series
	GPU	NVIDIA GeForce RTX 3060 Laptop
Computation Task	Purpose	Model Training, evaluation and deployment
Evaluation Metrics	Metrics	Precision, Recall, mAP@50, mAP@50–95

2.3 Dataset Preparation

2.3.1 Research Framework

Data acquisition was conducted using a Logitech full HD camera to capture image and video samples of the target objects. The collected data then underwent several preprocessing stages, including auto-orientation, contrast stretching, and image resizing to standardize input dimensions and enhance visual quality. To increase dataset diversity and robustness, various data augmentation techniques were applied, such as flipping, rotation, saturation adjustment, brightness variation, exposure modification, and blur noise injection. These strategies help simulate real-world variations and prevent model overfitting, enabling better generalization across different environmental conditions.

The augmented dataset was subsequently utilized to train and validate the YOLOv11 model. Model performance was evaluated using standard metrics, including mean Average Precision (mAP), precision, recall, and F1-score. Finally, a deployment analysis was conducted to assess the model's real-world applicability and operational reliability, as illustrated in Figure 2.

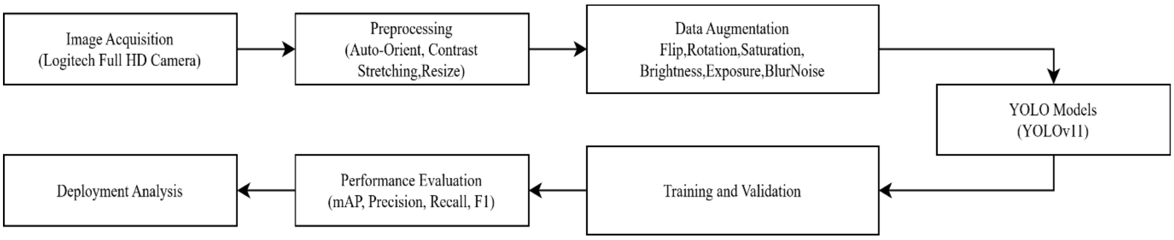


Figure 2. Research Framework pipeline

2.3.2 Dataset Description

The dataset consists of both image and video data captured using a camera, containing two primary object classes: drone and person, as illustrated in Figure 3(a) and Figure 3(b), respectively. Data collection was conducted under diverse environmental conditions, including variations in camera angles, object distances, and lighting scenarios, to accurately represent real-world operational settings. Each image was manually annotated using bounding boxes to precisely indicate the location of target objects.

The dataset comprises a total of 2082 images, which were systematically divided into training, validation, and testing subsets, as shown in Figure 4. This data partitioning strategy ensures an objective and measurable training and evaluation process, supporting reliable performance assessment and enhancing the generalization capability of the proposed YOLOv11 model.

(a) (b)

Figure 3. Object Class (a)Drone,(b)Person

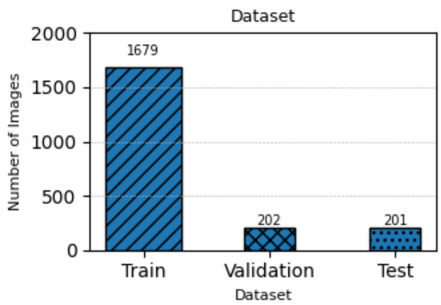


Figure 4. Dataset

2.3.3 Preprocessing and Augmentation

Preprocessing was applied to enhance data quality before being fed into the YOLOv11 model. The applied steps include automatic image orientation correction (Auto-Orient), image resizing through stretching to a standard resolution of 640 × 640 pixels, and adaptive contrast enhancement using contrast stretching to improve object visibility. These techniques enable the model to extract visual features more effectively, contributing to improved detection accuracy and system reliability, particularly during CPU-based deployment.

To increase dataset diversity and prevent overfitting, data augmentation techniques were employed, where each training sample generated three augmented outputs. The applied methods include flipping, random rotation (−5° to +5°), saturation adjustment (−20% to +20%), brightness and exposure variation (−25% to +25%), Gaussian blur up to 2.5 pixels, and random noise injection affecting up to 2.5% of image pixels. These augmentations simulate real-world conditions, enabling the YOLOv11 model to become more robust in detecting drones and humans in outdoor condition, as illustrated in Figure 5.

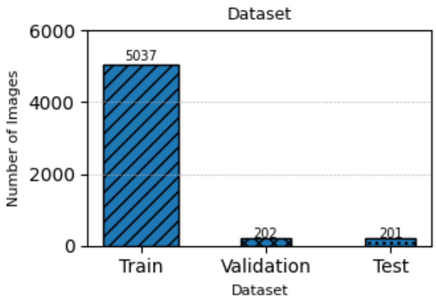


Figure 5. Distribution Augmentation Dataset

2.4 Evaluation Metrics

Detection performance in this study is evaluated using standard object detection metrics, including Precision, Recall, F1-score, Intersection over Union (IoU), mean Average Precision at an IoU threshold of 0.5 (mAP@0.5), mean Average Precision averaged over IoU thresholds from 0.5 to 0.95 (mAP@0.5:0.95), as well as inference efficiency metrics such as inference time and Frames Per Second (FPS). These metrics collectively assess both detection accuracy and real-time applicability.

2.4.1. Precision

Precision illustrates the ratio of accurately detected objects to the total number of detected objects, as defined in Equation (1). A high precision value indicates a diminished occurrence of false positives.

$$P = \frac{TP}{TP+FP} \tag{1}$$

Where *TP* and *FP* denote true positives and false positives, respectively.

2.4.2. Recall

Recall, as defined in Equation (2), illustrates the ratio of accurately detected objects to the total number of actual objects present in the image. A high recall indicates a diminished occurrence of false negatives.

$$R = \frac{TP}{TP+FN} \tag{2}$$

Where FN refers to false negatives.

2.4.3. F1-Score

The F1-score, as defined in Equation (3), provides a balanced evaluation of detection performance by computing the harmonic mean of precision and recall, especially when class distributions are imbalanced.

$$F1 = \frac{2 \cdot P \cdot R}{P+R} \tag{3}$$

2.4.4. Intersection over Union (IoU)

Intersection over Union (IoU), as defined in Equation (4), measures the degree of overlap between predicted and ground-truth bounding boxes, where higher values indicate better localization accuracy.

$$\text{IoU} = \frac{Area\ of\ overlap}{Area\ of\ Union} \tag{4}$$

Higher IoU values (e.g., IoU ≥ 0.5) indicate better localization performance.

2.4.5. Mean Average Precision (mAP)

Mean Average Precision (mAP) evaluates object detection performance by averaging precision–recall characteristics across object classes and IoU thresholds. The mAP@0.5, as defined in Equation (5) , represents the mean Average Precision across all classes at an IoU threshold of 0.5.

$$\mathrm{mAP}_{0.5} = \frac{1}{N}\sum_{c=1}^{N} AP_c(\mathrm{IoU} = 0.5) \tag{5}$$

where N denotes the number of object classes and $AP_{(c)}$ represents the Average Precision for class c.
To provide a stricter evaluation, mAP@0.5:0.95, defined in Equation (6), provides a more stringent evaluation by averaging mAP over multiple IoU thresholds from 0.5 to 0.95.

$$\mathrm{mAP}_{0.5:0.95} = \frac{1}{10}\sum_{t\in\{0.5,0.55,\ldots,0.95\}} \mathrm{mAP}_{\mathrm{IoU}=t} \tag{6}$$

2.4.6. Inference Time and Frames Per Second (FPS)

Inference efficiency is assessed using inference time per frame and Frames Per Second (FPS). The inference time per frame, defined in Equation (7), represents the average processing duration for a single frame, whereas FPS, defined in Equation (8)(9), measures the number of frames processed per second, indicating the model's suitability for real-time deployment.

$$t_{infer} = \frac{T_{Infer}}{Number\ of\ frames} \tag{7}$$

$$\mathrm{FPS} = \frac{1}{t_{\mathrm{inference}}} \tag{8}$$

$$\mathrm{FPS} = \min(\frac{1}{t_{\mathrm{inference}}}, FPS_{\mathrm{max}}) \tag{9}$$

These metrics are particularly important for real-time applications such as drone-assisted oil palm plantation monitoring.

2.5 Deployment-Oriented Evaluation Setup

Figure 6.Show deployment setup is designed to evaluate the real-world performance of the proposed YOLOv11-based detection system under practical operating conditions. The system pipeline begins with real-time video acquisition from a camera, which continuously captures visual data in the operational environment. The incoming frames are directly forwarded to the YOLOv11 model, where object detection is performed to identify drones and persons in each frame.

The inference process is executed on a CPU-based computing platform to simulate low-resource deployment scenarios. This setup aims to assess the feasibility of running the detection model on edge devices or embedded systems without GPU acceleration. Performance metrics such as inference time, frame processing rate (FPS), CPU utilization, and detection accuracy are monitored to evaluate system efficiency and stability.

After inference, the detection results, including bounding boxes and class labels, are visualized and recorded for further analysis. The final deployment stage demonstrates the system's capability to operate in real-time surveillance scenarios, validating its readiness for field implementation. This evaluation setup ensures that the proposed system is not only accurate but also computationally efficient and suitable for real-world deployment.

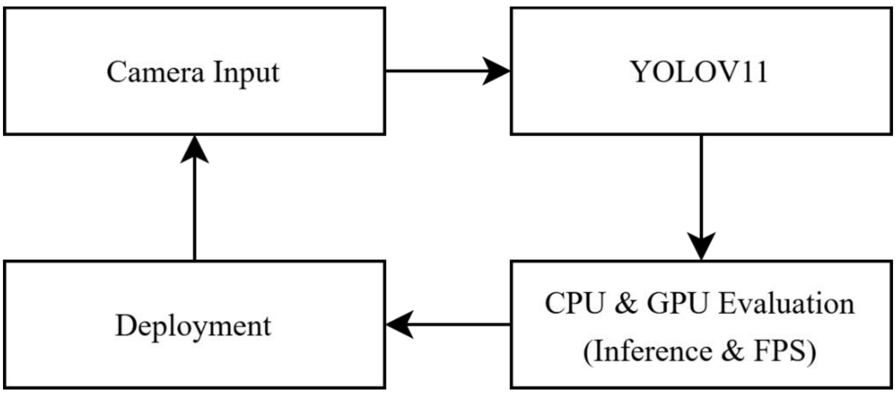


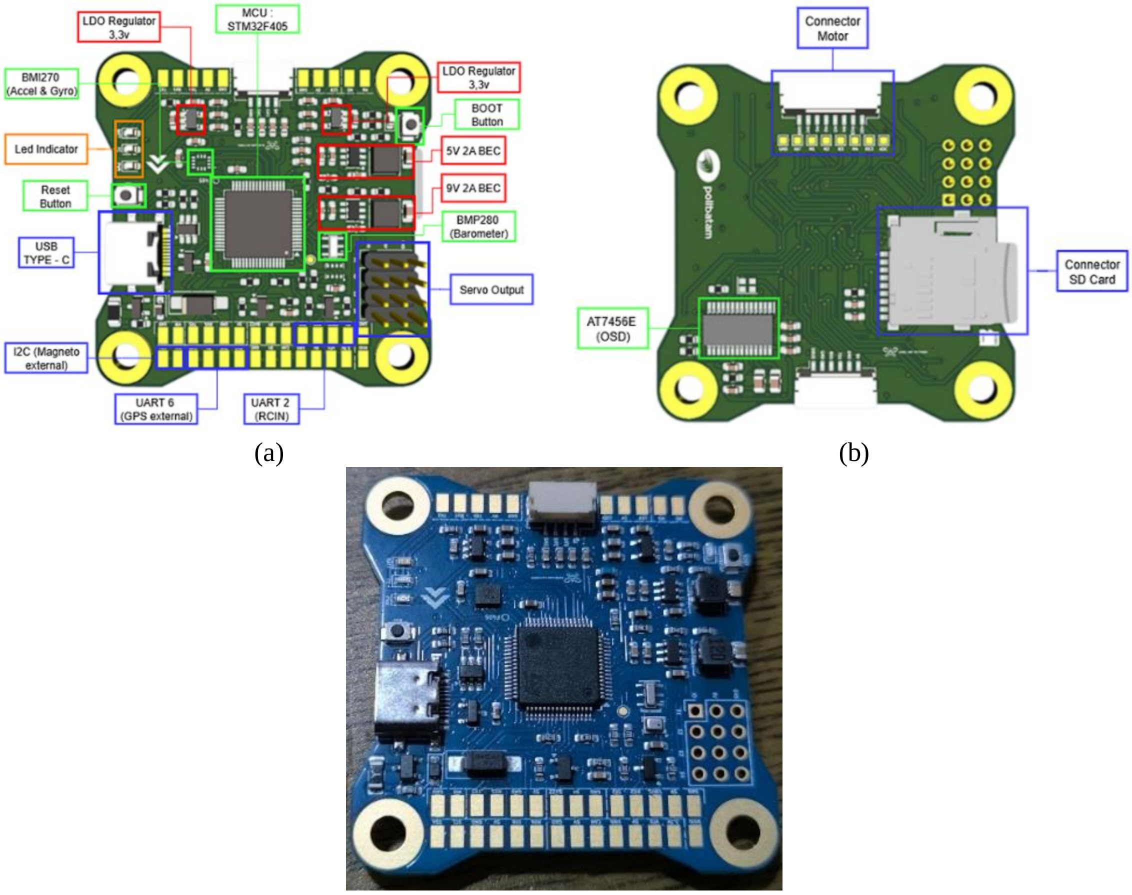
Figure 6. Deployment Pipeline

2.6 Drone Framework
2.6.1 Custom Flight Controller PCB Design.

This study utilizes a custom-designed flight controller PCB developed specifically for quadcopter UAV applications. The hardware was designed using professional PCB design software to ensure optimal component placement, signal integrity, and a compact layout. The flight controller adopts a Pixhawk-class architecture to support real-time flight stabilization and navigation.

The PCB integrates multiple interfaces, including IMU connections (I2C/SPI), PWM motor outputs, UART/USB communication ports, and power management circuits. This custom hardware design increases system flexibility, reduces dependency on commercial flight controllers, and lowers development costs. Furthermore, the PCB supports scalability and modularity for future upgrades through expansion headers that enable the integration of peripherals such as GPS modules and telemetry radios.

Figure 7 illustrates the physical realization of the proposed PCB. Figure 7(a) shows the top side of the PCB, which contains the main microcontroller unit (STM32), voltage regulators, clock crystal, and signal conditioning components responsible for sensor fusion and control processing. Figure 7(b) presents the bottom side of the PCB, where decoupling capacitors, communication traces, and grounding layers are implemented to enhance signal stability and electromagnetic compatibility, while Figure 7(c) illustrates the realization visualization of the PCB design used.



(a) (b)
(c)

Figure 7. Costum Flight 3D Design PCB flight controller: (a) Top side, (b) Bottom side,(c) Physical realization of the designed PCB

2.6.2 UAV Overall Configuration

The proposed UAV-based drone–human differentiation system consists of three main layers: vision acquisition, processing unit, and flight control subsystem. A Logitech Full HD USB camera is connected to an onboard laptop through a wired USB active repeater cable interface to continuously capture real-time video frames. These frames are processed by the laptop, where the YOLO11s model performs object detection to classify drones and humans in real time. The detection results are transmitted to the UAV system via a telemetry communication link, enabling real-time data exchange between the processing unit and the custom flight controller.

The custom flight controller receives the detection information and integrates it with onboard sensor data, including inputs from the GPS module for navigation, a radio receiver for manual control commands, and electronic speed controllers (ESCs) for propulsion management. Based on system logic and control commands, the flight controller generates PWM signals to regulate motor speed and maintain UAV stability. This system architecture supports a gradual transition from offline video analysis to real-time deployment, ensuring system reliability prior to field operation, as illustrated in Figure 8.

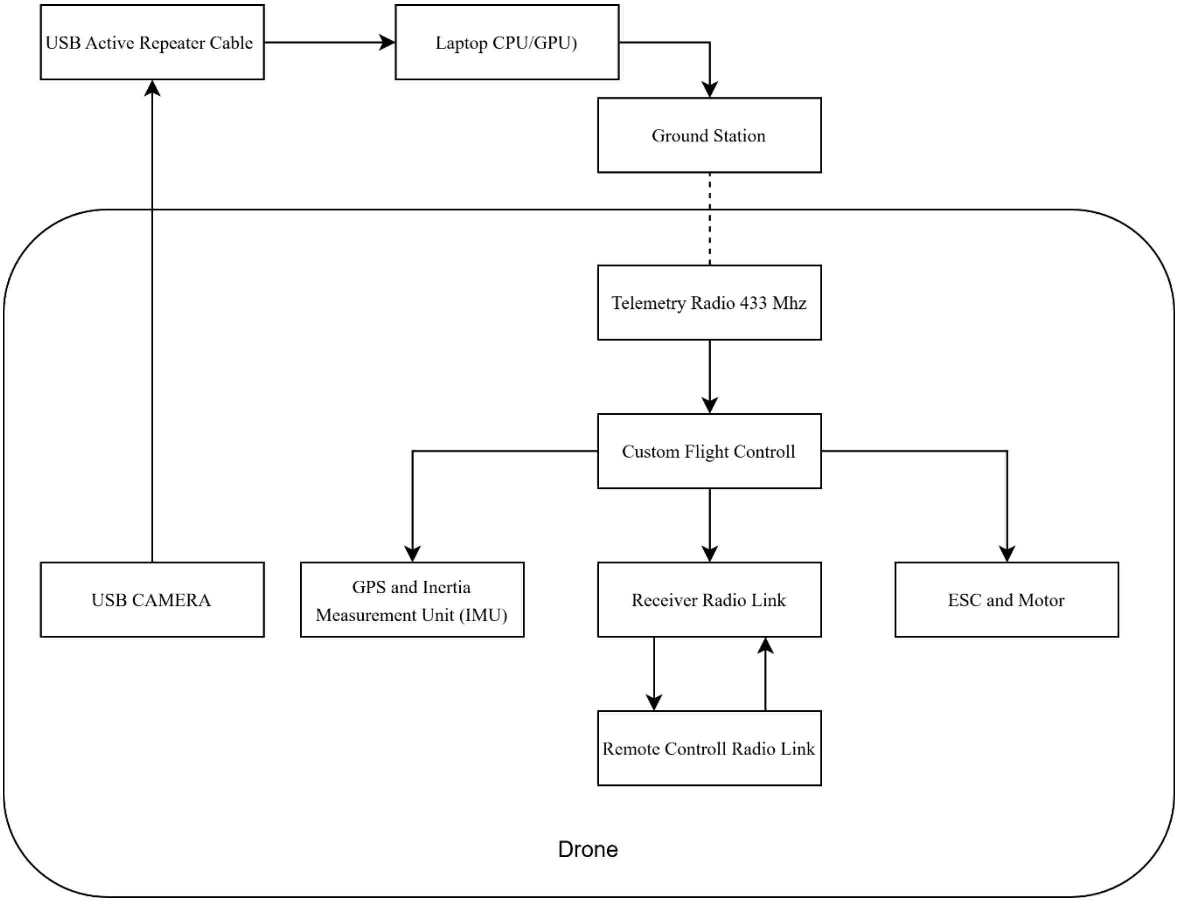


Figure 8. System UAV-based drone–human differentiation overall configuration

2.6.3 Drone Vehicle Used

Figure 9(a) shows the integration of the custom-designed PCB within the drone system, demonstrating its placement and connectivity with other onboard components. Meanwhile, Figure 9(b) illustrates the camera module that has been successfully integrated into the UAV platform, enabling real-time visual data acquisition for monitoring and detection purposes.

(a) (b)

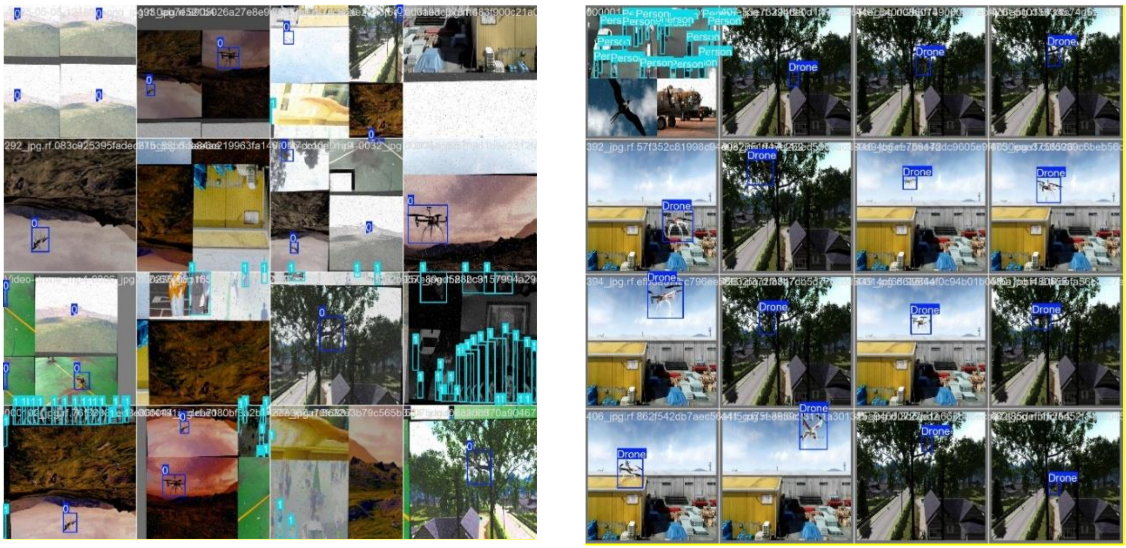
Figure 9. Drone : (a)Integration Flight Controll,(b)Integration camera

3. RESULTS AND DISCUSSION

3.1 Training Batch and Validation Batch

Figure 10(a) illustrates a sample training batch consisting of annotated images with various drone and person instances under different scales, orientations, and background conditions. This diversity reflects the effectiveness of the applied data augmentation techniques, such as flipping, rotation, brightness adjustment, and noise injection, which enhance model generalization.

Figure 10(b) presents the validation batch used to evaluate model performance on unseen data. These images are excluded from the training process, ensuring an unbiased assessment of detection accuracy. The validation samples represent real-world scenarios with variations in illumination, occlusion, and object distance.



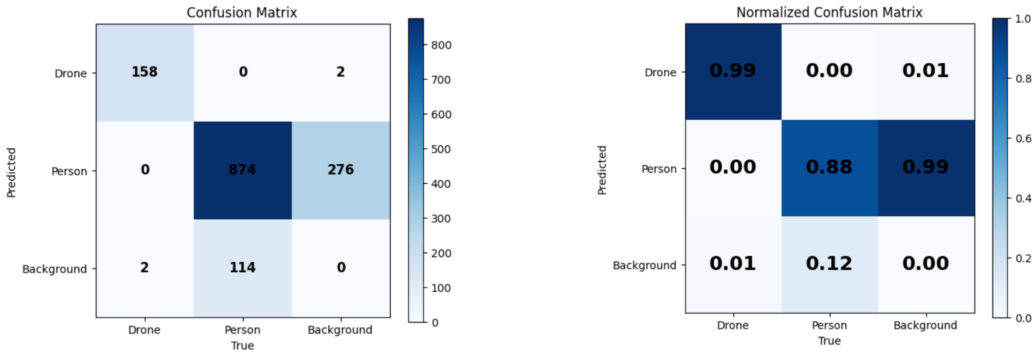
(a) (b)

Figure 10. (a)Train Batch,(b)Val Batch

3.2 Overall Training Performance

Figure 11(a) shows the raw confusion matrix of the proposed model. The drone class achieves 158 correct detections with only 2 misclassifications as background, indicating strong detection reliability. The person class records 874 correct predictions; however, 276 samples are misclassified as background, showing that the model still faces challenges in distinguishing persons from complex backgrounds. Meanwhile, the background class is often confused with the person class, with 114 samples incorrectly classified as persons.

Figure 11(b) presents the normalized confusion matrix, which further clarifies class-wise performance. The drone class achieves a high accuracy of 0.99, demonstrating excellent detection robustness. The person class reaches an accuracy of 0.88, while approximately 0.12 of person samples are misclassified as background. This indicates difficulties in handling occlusion and cluttered environments. Overall, these results confirm that the proposed YOLO11s model provides strong generalization capability while highlighting areas for improvement in complex scenes.



(a) (b)

Figure 11. (a)Confusion Matrix and (b)Normalized Confusion Matrix

Figure 12 presents the training and validation performance curves of the proposed YOLO11s model across 100 epochs. The box loss, classification loss, and distribution focal loss for both training and validation sets exhibit a consistent decreasing trend, indicating stable learning behavior and effective convergence. This demonstrates that the model successfully minimizes prediction errors during training without significant overfitting. Additionally, precision and recall curves show steady improvement and stabilize above 0.90, confirming reliable object detection performance. The mAP@50 and mAP@50–95 metrics also increase progressively and converge at approximately 0.93 and 0.72, respectively, reflecting strong localization accuracy across different IoU thresholds.

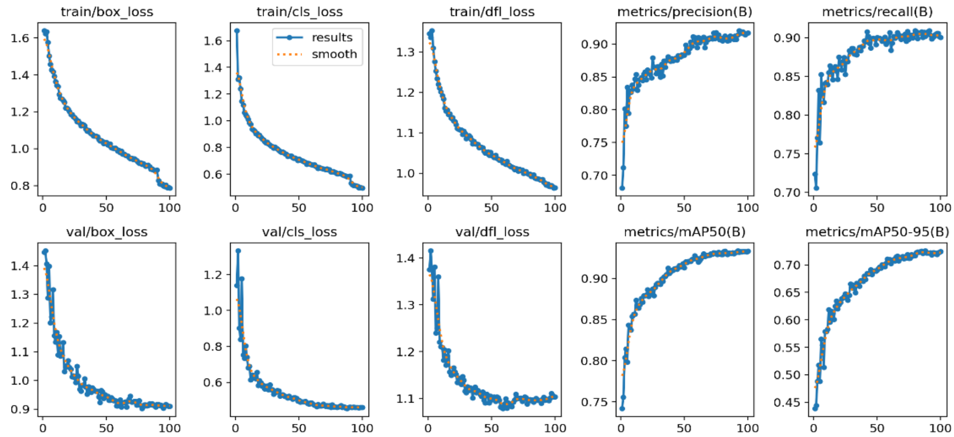


Figure 12. Deployment Result

Table 2 summarizes the quantitative performance of the trained model. The proposed approach achieves a precision of 0.912 and a recall of 0.903, indicating a strong balance between correct detection and false alarm reduction. The resulting F1-score of 0.907 further confirms the robustness of the model in handling class imbalance. Moreover, the high mAP@50 value of 0.932 demonstrates excellent detection accuracy, while the mAP@50–95 score of 0.727 highlights the model's consistency under more stringent evaluation criteria. Overall, these results validate the effectiveness and reliability of the proposed YOLO11s-based detection system for drone–human differentiation in real-world applications.

Table 2. Training Performance

Metric	Value
Precision	0.912
Recall	0.903
F1-Score	0.907
mAP@50	0.932
mAP@50–95	0.727

3.3 Evaluation Metrcis Calculation

The F1-score is used to evaluate the balance between precision and recall based on the results presented in Table 2. Using the obtained precision value of 0.912 and recall value of 0.903, the calculated F1-score reaches 0.907, as summarized in Table 3. This result indicates that the proposed detection model achieves a well-balanced performance, effectively maintaining high detection accuracy while minimizing false predictions. The obtained F1-score further confirms the reliability and robustness of the model for drone–human differentiation tasks.

Table 3. F1-Score Evaluation

F1-Score
$F1 = \frac{2 \cdot P \cdot R}{P + R}$ $F1 = \frac{2 \cdot 0.912 \cdot 0.903}{0.912 + 0.903}$ $F1 = \frac{1.646}{1.815}$ $F1 = 0.907$ F1-Score = 0.907

3.4 Video Detection Validation Test

Figure 13 presents the real-time validation performance of the proposed YOLOv11s model in an outdoor environment. The system successfully detects multiple objects, including a drone and two human subjects, with high confidence scores. The drone object is detected with a confidence level of 0.82, while the person class achieves confidence values of 0.84 and 0.89, respectively. These results indicate the robustness of the model in simultaneously identifying aerial and human targets under natural lighting conditions and complex backgrounds. Furthermore, the consistent detection accuracy demonstrates the model's suitability for real-world surveillance and security applications.

Figure 13. Video Validation Testing

3.5 Real Time Detection

Figure 14 presents the real-time deployment results of the proposed drone–human detection system evaluated on both CPU and GPU platforms. The left frame corresponds to CPU-based inference, while the right frame shows GPU-accelerated inference. The system successfully detects and classifies drones and humans in an outdoor environment, achieving confidence scores of 0.86 for drone detection and 0.71 for human detection.

Quantitative performance metrics are overlaid on each frame. The CPU implementation achieves an average processing speed of 5.2 FPS with an inference latency of 191.8 ms per frame, which is insufficient for high-speed real-time applications. In contrast, GPU acceleration significantly improves performance, reaching 12.4 FPS with a reduced inference time of 81.0 ms, representing a latency reduction of approximately 57.8% compared to the CPU implementation.

These results clearly demonstrate the effectiveness of hardware acceleration in enhancing real-time detection capability. The proposed system maintains detection accuracy while achieving substantially lower latency and higher throughput on GPU platforms. This performance improvement confirms the feasibility of deploying the proposed model in real-world surveillance and aerial security scenarios, where fast response time and reliable object discrimination are critical.

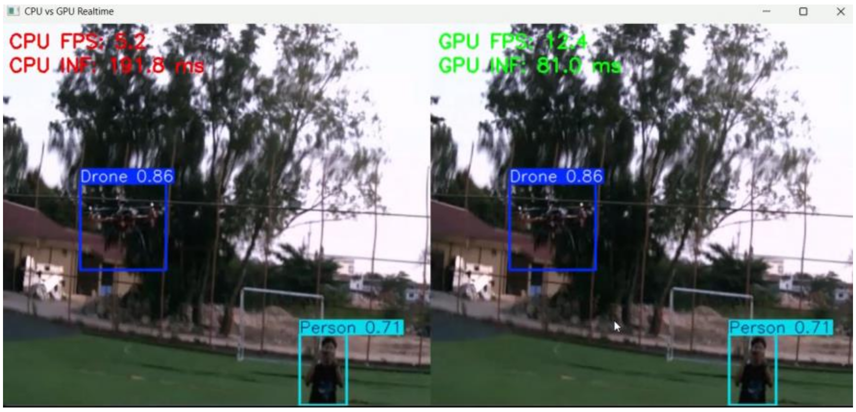
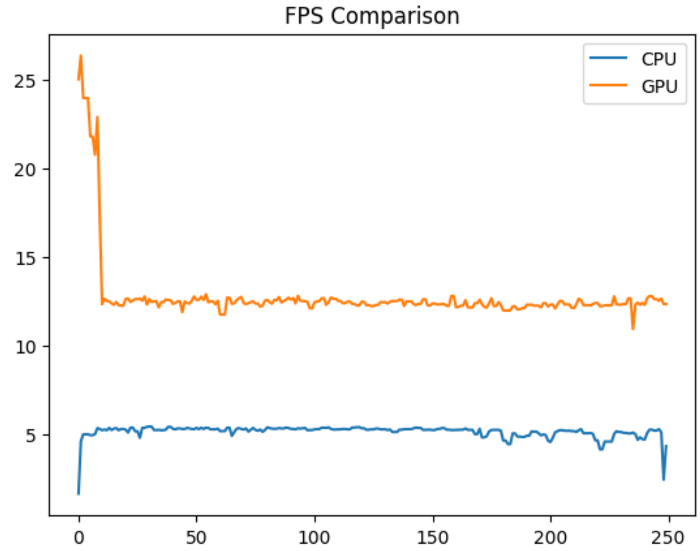


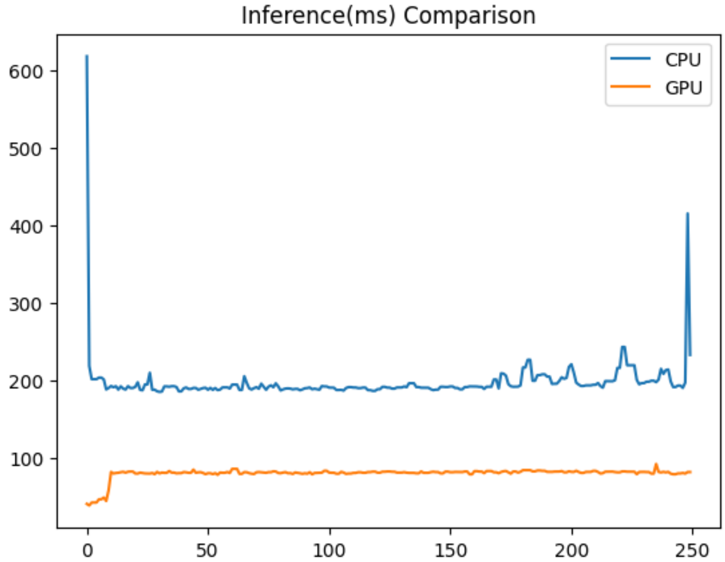
Figure 14. Real-Time Deployment Testing

Figure 15(a) shows the real-time FPS comparison between CPU and GPU implementations. The GPU consistently achieves a higher and more stable frame rate, maintaining an average of approximately 12–13 FPS, with an initial peak above 25 FPS during warm-up. In contrast, the CPU performance remains significantly lower, stabilizing around 5 FPS throughout the testing period.These results indicate that GPU acceleration provides more than 2× improvement in throughput compared to CPU execution. The stable FPS trend on GPU demonstrates its suitability for real-time deployment, while the CPU performance is limited by computational constraints, making it less optimal for high-speed applications.

Figure 15(b) illustrates the inference time comparison between CPU and GPU platforms. The CPU exhibits a significantly higher latency, with an average inference time of approximately 190–210 ms, including several spikes exceeding 400 ms, indicating processing instability. Conversely, the GPU maintains a consistently low inference time of approximately 80–85 ms with minimal fluctuation.The GPU implementation achieves a latency reduction of over 55% compared to the CPU, confirming the effectiveness of hardware acceleration. This substantial decrease in inference time directly contributes to the higher FPS observed in Figure 15(a), reinforcing the feasibility of GPU-based deployment for real-time drone–human detection systems.



(a)



(b)

Figure 15. Performance Comparison Between CPU and GPU

3.6 Real Time Detection CPU & GPU Calculation

The real-time performance evaluation results presented in Figure 14 are quantitatively analyzed using Eq. (7),(8). The CPU-based implementation records an inference time of 191.8 ms, which corresponds to 0.1918 s. By substituting this value into Eq.(7),(8), the calculated FPS is shown in Table 4 :

Table 4. Inference and FPS Evaluation

Device	Inference (ms)	FPS	FPS Result	FPS Result Figure 14
CPU	191.8	$t_{infer} = \frac{T_{Infer}}{Number\ of\ frames}$ $t^{CPU}_{Inference} = \frac{191.8}{1000} = 0.1918\ s$ Convert to seconds $\text{FPS}_{CPU} = \frac{1}{t_{\text{inference}}}$ $\text{FPS}_{CPU} = \frac{1}{0.1918}$ $\text{FPS}_{CPU} = 5.21$	5.21 FPS	5.21 FPS
GPU	81.0	$t_{infer} = \frac{T_{Infer}}{Number\ of\ frames}$ $t^{GPU}_{Inference} = \frac{81}{1000} = 0.081\ s$ Convert to seconds $\text{FPS}_{GPU} = \frac{1}{t_{\text{inference}}}$ $\text{FPS}_{GPU} = \frac{1}{0.081}$ $\text{FPS}_{GPU} = 12.4$	12.4 FPS	12.4 FPS

The calculated FPS values are identical to those displayed in Figure 14, confirming the accuracy of the real-time performance measurement method. These results demonstrate that GPU acceleration significantly improves processing speed compared to CPU execution, validating its effectiveness for real-time deployment scenarios.

4. CONCLUSION

This study presents a real-time vision-based framework for drone–human differentiation using a lightweight YOLO11s model. In accordance with the objectives stated in the Introduction, the proposed approach demonstrates strong detection performance by integrating hybrid datasets and structured data augmentation strategies. Experimental results show that the model achieves a precision of 0.912, recall of 0.903, and an F1-score of 0.907, confirming its robustness and generalization capability under complex outdoor environments. These findings validate the effectiveness of the proposed system for aerial security applications requiring accurate and reliable object detection.

Furthermore, real-time deployment experiments on heterogeneous hardware platforms confirm the system's practical feasibility. The GPU-based implementation significantly outperforms CPU execution, achieving 12.4 FPS with an inference time of 81.0 ms, compared to 5.2 FPS and 191.8 ms on CPU. This performance gain highlights the importance of hardware acceleration for latency-sensitive surveillance systems. Future research will focus on extending the framework to multi-class detection, edge-device deployment, and sensor fusion to further enhance system scalability and robustness in real-world applications.

ACKNOWLEDGMENTS

The authors gratefully acknowledge Barelang Robotics and Artificial Intelligence (BRAIL) and the Department of Electrical Engineering, Batam State Polytechnic, for providing the facilities and financial support for this research. The authors also thank the Robotics Research Team at BRAIL for their technical assistance and valuable contributions. Special thanks are extended to Mr. Ahmad Riyad Firdaus, Chief Advisor, for his guidance and supervision throughout this study. Finally, the authors express their sincere appreciation to their families for their continuous support and encouragement.

FUNDING INFORMATION

This research was supported by internal funding and institutional support from Barelang Robotics and Artificial Intelligence (BRAIL) and the Department of Electrical Engineering, Batam State Polytechnic.

AUTHOR CONTRIBUTIONS STATEMENT

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Ezha Tri Saputra	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓			
Ahmad Riyad Firdaus	✓	✓		✓	✓	✓			✓	✓		✓	✓	✓
Daipansyah Arya S.			✓			✓								
Hendawan Soebhakti						✓						✓		
Emelia Rosari Siregar						✓						✓		

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITYs

The data that support the findings of this study are openly available on Roboflow Universe. The primary dataset used in this research was collected and curated by the authors and is available at

- The data that support the findings of this study will be available in https://universe.roboflow.com/ezha-tri-saputra/droneperson-oasdb/dataset/1
- The data that support the findings of this study are openly available in https://universe.roboflow.com/vasu12360-gmail-com/people-detection-a5s5p

REFERENCES

[1] H. Shakhatreh *et al.*, "Unmanned Aerial Vehicles: A Survey on Civil Applications and Key Research Challenges," Apr. 2018, doi: 10.1109/ACCESS.2019.2909530.

[2] "Otto, A., Agatz, N., Campbell, J., Golden, B., & Pesch, E. (2018). Optimization approaches for civil applications of unmanned aerial vehicles (UAVs) or aerial drones: A survey. Networks, 72(4), 411-458. https://doi.org/10.1002/net.21818".

[3] A. Famili, A. Stavrou, H. Wang, J. M. Park, and R. Gerdes, "Securing Your Airspace: Detection of Drones Trespassing Protected Areas," Apr. 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/s24072028.

[4] A. Names and S. Mohammed Zeki Abd Alsammed, "23 Drone Management System for Hunting Unauthorized Drones."

[5] G. Golcarenarenji, I. Martinez-Alpiste, Q. Wang, and J. M. Alcaraz-Calero, "Efficient Real-Time Human Detection Using Unmanned Aerial Vehicles Optical Imagery," *Int. J. Remote Sens.*, vol. 42, no. 7, pp. 2440–2462, Apr. 2021, doi: 10.1080/01431161.2020.1862435;WEBSITE:WEBSITE:TFOPB;PAGEGROUP:STRING:PUBLICATION.

[6] D. Simões, H. Oliveira, S. Marsico, J. Souza, and L. Aparecido Barbosa, "Human detection in UAV imagery using deep learning: a review," *Neural Comput. Appl.*, vol. 37, pp. 18109–18150, Jan. 2025, doi: 10.1007/s00521-025-11446-5.

[7] X. Wang, Z. Huang, B. Liao, L. Huang, Y. Gong, and C. Huang, "Real-Time and Accurate Object Detection in Compressed Video by Long Short-term Feature Aggregation," Mar. 2021, doi: 10.1016/j.cviu.2021.103188.

[8] H. R. Sinulingga and S. G. Kong, "Key-Frame Extraction for Reducing Human Effort in Object Detection Training for Video Surveillance," *Electronics (Switzerland)*, vol. 12, no. 13, Jul. 2023, doi: 10.3390/electronics12132956.

[9] G. Tang, J. Ni, Y. Zhao, Y. Gu, and W. Cao, "A Survey of Object Detection for UAVs Based on Deep Learning," *Remote Sens. (Basel).*, vol. 16, no. 1, Jan. 2024, doi: 10.3390/rs16010149.

[10] Z. Yuan *et al.*, "Small Object Detection in UAV Remote Sensing Images Based on Intra-Group Multi-Scale Fusion Attention and Adaptive Weighted Feature Fusion Mechanism," *Remote Sens. (Basel).*, vol. 16, no. 22, Nov. 2024, doi: 10.3390/rs16224265.

[11] Z. Liu, P. An, Y. Yang, S. Qiu, Q. Liu, and X. Xu, "Vision-Based Drone Detection in Complex Environments: A Survey," Nov. 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/drones8110643.

[12] G. Gutierrez, J. P. Llerena, L. Usero, and M. A. Patricio, "A Comparative Study of Convolutional Neural Network and Transformer Architectures for Drone Detection in Thermal Images," *Applied Sciences (Switzerland)*, vol. 15, no. 1, Jan. 2025, doi: 10.3390/app15010109.

[13] J. Zhang, "Towards a High-Performance Object Detector: Insights from Drone Detection Using ViT and CNN-based Deep Learning Models."

[14] H. Lee *et al.*, "CNN-Based UAV Detection and Classification Using Sensor Fusion," *IEEE Access*, vol. 11, pp. 68791–68808, 2023, doi: 10.1109/ACCESS.2023.3293124.

[15] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," Jul. 2022, [Online]. Available: http://arxiv.org/abs/2207.02696

[16] M. A. R. Alif and M. Hussain, "YOLOv1 to YOLOv10: A comprehensive review of YOLO variants and their application in the agricultural domain," Jun. 2024, [Online]. Available: http://arxiv.org/abs/2406.10139

[17] J. Terven and D. Cordova-Esparza, "A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS," Feb. 2024, doi: 10.3390/make5040083.

[18] J. Bento, T. Paixão, and A. B. Alvarez, "Performance Evaluation of YOLOv8, YOLOv9, YOLOv10, and YOLOv11 for Stamp Detection in Scanned Documents," *Applied Sciences (Switzerland)*, vol. 15, no. 6, Mar. 2025, doi: 10.3390/app15063154.

[19] Z. Liu, P. An, Y. Yang, S. Qiu, Q. Liu, and X. Xu, "Vision-Based Drone Detection in Complex Environments: A Survey," Nov. 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/drones8110643.

[20] G. Tang, J. Ni, Y. Zhao, Y. Gu, and W. Cao, "A Survey of Object Detection for UAVs Based on Deep Learning," *Remote Sens. (Basel).*, vol. 16, no. 1, Jan. 2024, doi: 10.3390/rs16010149.

[21] J. Yang, S. Liu, Z. Li, X. Li, and J. Sun, "Real-time Object Detection for Streaming Perception."

[22] S. Xie, G. Deng, B. Lin, W. Jing, Y. Li, and X. Zhao, "Real-Time Object Detection from UAV Inspection Videos by Combining YOLOv5s and DeepStream," *Sensors*, vol. 24, no. 12, Jun. 2024, doi: 10.3390/s24123862.

[23] P. Wu, C. Pan, Y. Yan, G. Pang, P. Wang, and Y. Zhang, "Deep Learning for Video Anomaly Detection: A Review," Sep. 2024, [Online]. Available: http://arxiv.org/abs/2409.05383

[24] N. Alsharabi, "Real-Time Object Detection Overview: Advancements, Challenges, and Applications," 2023.

[25] K. Sharma, T. K. Sahu, P. Bhatt, A. K. Singh, and M. Bhatia, "Real Time Video Object Detection Using Deep Learning," *2023 International Conference on Advances in Computation, Communication and Information Technology, ICAICCIT 2023*, pp. 601–606, 2023, doi: 10.1109/ICAICCIT60255.2023.10466207.
[26] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *J. Big Data*, vol. 6, no. 1, Dec. 2019, doi: 10.1186/s40537-019-0197-0.
[27] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," *Information (Switzerland)*, vol. 11, no. 2, Feb. 2020, doi: 10.3390/info11020125.
[28] M. Sarwar Awan, S. Azhar Ali Zaidi, and J. Mir, "UETT4K Anti-UAV: A Large Scale 4K Benchmark Dataset for Vision-Based Drone Detection in High-Resolution Imagery," *IEEE Access*, vol. 13, pp. 73553–73568, 2025, doi: 10.1109/ACCESS.2025.3562890.
[29] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," May 2016, [Online]. Available: http://arxiv.org/abs/1506.02640
[30] M. L. Ali and Z. Zhang, "The YOLO Framework: A Comprehensive Review of Evolution, Applications, and Benchmarks in Object Detection," Dec. 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/computers13120336.

BIOGRAPHIES OF AUTHORS

Ezha Tri Saputra is a Robotics Engineering student at Batam State Polytechnic, Indonesia, and a member of the Flying Robot Division since 2022. His research interests include aerial robotics, UAV systems, flight control and embedded systems, autonomous navigation, and AI-based computer vision for drone applications. He has experience developing quadcopter and fixed-wing platforms, PCB-based flight controllers, sensor integration, and real-time control, as well as implementing YOLO on edge devices for precision agriculture,aerial security and monitoring. He can be contacted at email:Trisaputraezha@gmail.com

Ir. Ahmad Riyad Firdaus, Ph.D. is a faculty member of the Electrical Engineering Department and currently serves as Vice Director for Academic Affairs at Politeknik Negeri Batam, Indonesia. He earned his Ph.D. in Automatic Control and Systems Engineering from the University of Sheffield, United Kingdom, and completed his bachelor's and master's degrees at Institut Teknologi Bandung (ITB), Indonesia. His research focuses on nonlinear and robust control systems, soft computing, robotics, UAVs, and embedded systems. With over two decades of experience in higher education, Dr. Firdaus has been instrumental in advancing Project-Based Learning (PBL) and promoting the CDIO framework in Indonesia's vocational higher education sector. He has also contributed to expanding the Indonesian Accreditation Board for Engineering Education (IABEE) accreditation scope in line with the Sydney and Dublin Accords and serves as an IABEE Program Evaluator. His dedication to curriculum innovation, international accreditation, and educational leadership has earned him multiple awards for excellence and impact in vocational higher education.

Daipansyah Arya Saputra is a Robotics Engineering student at Batam State Polytechnic, Indonesia, and has been an active member of the Flying Robot Division since 2022. He was born in Pagaralam on October 3, 2003. Currently, he is pursuing a degree in Electrical Engineering with a specialization in Robotics Engineering Technology.
His research interests focus on aerial robotics, UAV systems, flight control, and embedded systems. He has been actively involved in various academic and research projects related to unmanned aerial vehicles, particularly in the development of flight controllers and intelligent navigation systems. Through his participation in the Flying Robot Division, he has contributed to the design, implementation, and testing of drone platforms for research and competition purposes.He can be contacted via email: arya45927@gmail.com.

	**Hendawan Soebhakti, S.T., M.T.,** is an assistant professor in Robotics Engineering Technology, **Politeknik Negeri Batam (Polibatam)**. He serves as the Expert Staff to the Vice Director for Academic Affairs at Polibatam and is the Secondary Contact within the CDIO Worldwide network. He earned his Master's Degree in Electronics Engineering with a concentration in Robotics Technology from Institut Teknologi Sepuluh Nopember (ITS) in 2012. Since 2015, he has consistently driven educational innovation by developing and implementing Project-Based Learning (PBL) methods. He spearheaded the establishment of the Robotics Engineering Technology program, applied the CDIO framework starting in 2021, and led Polibatam to become an official CDIO member in 2022. To ensure long-term institutional sustainability, he initiated the CDIO Taskforce at Polibatam to oversee and coordinate CDIO implementation across study programs. He is a prominent national speaker on Project-Based Learning (PBL), Outcome-Based Education (OBE), and CDIO. His pedagogical interests focus on robotics development for STEM education, while his research concentrates on mobile robot technologies, particularly Automated Guided Vehicles (AGV) and Autonomous Mobile Robots (AMR). He also serves as the Chairperson of the Instructional Techniques Basic Skills Enhancement Training (PEKERTI).
	**Emelia Rosari Siregar** is a lecturer in Language, Linguistics, and Communication at Politeknik Negeri Batam (Polibatam), Indonesia. She specializes in English for Professional Skills and Written/Spoken Communication, teaching courses such as English for Written Communication, English for the Workplace, and Technical English: Presentation within the Robotics Engineering Technology (RET) study program, where she also serves as an academic advisor. Her research and professional interests center on English for Specific Purposes (ESP), with a focus on building technical communication competencies among engineering students. She is actively involved in curriculum validation and program accreditation, and has contributed to community engagement initiatives connecting robotics education with industry and international partners.