Jurnal Teknik Informatika C.I.T Medicom, 18 (2) (2026) 133-150

Published by: Institute of Computer Science (IOCS)

Jurnal Teknik Informatika C.I.T Medicom

Journal homepage: www.medikom.iocspublisher.org

Real-time human detection on FPV drones using YOLOv11 and ESP-NOW

Aria Kusumah Sastradinata[1], Bagus Hendra Saputra[2], Rifky Adishatya[3], Gumayang Fitri Annisa[4], Lusy Amelia[5], Belinda zhafira[6], Mukhamad Ayx T Zus Rizal Tofa[7]

[1,2,3,4,5,6,7] Republic of Indonesia Defense University, Bogor, Indonesia

Article Info

Article history

Received : Apr 30, 2026
Revised : May 26, 2026
Accepted : May 31, 2026

Keywords:

ESP-NOW;
FPV Drone;
Human Detection;
Real-Time Detection;
Temporal Validation.

Abstract

Conventional aerial surveillance systems still rely heavily on human operators, which may lead to visual fatigue, limited monitoring coverage, and delayed responses during security patrol operations. This study proposes a real-time human detection system for FPV drone surveillance using the YOLOv11 object detection model integrated with ESP-NOW wireless communication. The proposed system incorporates temporal validation and human-in-the-loop confirmation to improve detection reliability and maintain operator control during response activation. Experimental evaluations were conducted under morning, afternoon, and evening conditions. The proposed system achieved average confidence values of 81.25%, 78.38%, and 79.88%, with detection success rates of 71.13%, 75.94%, and 78.03%, respectively. Furthermore, the ESP-NOW communication subsystem successfully transmitted activation signals with delays ranging from 7 ms to 53 ms and maintained stable communication over distances up to 300 m. The main contribution of this research lies in the integration of YOLOv11, temporal validation, human-in-the-loop confirmation, and ESP-NOW communication into a single UAV surveillance framework, enabling reliable real-time human detection while preserving human supervision in operational decision-making.

Corresponding Author:

Aria Kusumah Sastradinata
Faculty of Defense Engineering and Technology
Republic Indonesia Defense University
Kawasan IPSC Sentul, Sukahati, Kec. Citeureup, Bogor, Jawa Barat, 16810, Indonesia
Email: ariakusumah2005@gmail.com

1. Introduction

Modern aerial surveillance systems continue to depend largely on human operators to observe and interpret visual information manually. This reliance often creates several operational challenges, such as visual exhaustion, misinterpretation of objects, limited observation coverage, and reduced efficiency when monitoring extensive areas under constantly changing environmental conditions. In security patrols, maritime monitoring, and traffic supervision, inaccurate object identification may contribute to slower response times and reduced effectiveness in operational decision making.

Comparable conditions are also encountered at Republic Indonesia Defense University, where the campus environment consists of wide operational areas, multi-story buildings, and high mobility of personnel and vehicles. These conditions require a surveillance system that is both adaptive and efficient. Based on interviews with campus security personnel, conventional patrol activities still

experience major obstacles when monitoring broad areas, particularly the upper sections of faculty and administrative buildings that cannot be observed optimally from ground level. In addition, long patrol durations and exposure to outdoor weather conditions can reduce operator concentration and visual accuracy. Although CCTV systems are available, their coverage is primarily limited to indoor areas, leaving many outdoor locations insufficiently monitored. These limitations highlight the importance of developing an aerial monitoring system capable of supporting patrol operations more effectively and systematically.

The emergence of Unmanned Aerial Vehicles (UAVs) has provided a practical solution for expanding monitoring coverage and improving operational mobility. Nevertheless, many UAV-based surveillance systems still function merely as passive observation tools that transmit live aerial video streams to operators. In most cases, object interpretation and threat assessment are still conducted manually, which means the possibility of human error remains significant. For this reason, integrating artificial intelligence into UAV surveillance platforms has become increasingly necessary to enhance monitoring capability in complex and dynamic environments.

Recent advances in deep learning have significantly improved object detection technology by enabling systems to identify objects and localize them simultaneously within a single inference process [1]. Convolutional Neural Network (CNN)-based approaches have demonstrated strong performance in processing aerial imagery and improving the detection accuracy of small-scale objects [2], [3]. Among current object detection models, the YOLO (You Only Look Once) family is widely recognized for combining fast inference speed with high detection accuracy [4], [5]. The latest generation, YOLOv11, introduces improvements in computational efficiency and model generalization, making it suitable for real-time aerial surveillance applications [6].

This research focuses on the implementation of YOLOv11 for detecting humans from aerial imagery captured using a DJI FPV drone. Human detection plays an important role in modern surveillance and patrol systems because human presence often becomes the primary indicator in security monitoring activities [7]. However, relying solely on visual detection is still insufficient for real operational scenarios because detection outputs require a controlled decision-making mechanism to avoid unnecessary responses and incorrect system activation.

In real-time video processing, temporary detections may occur because of rapid object movement, pose changes, and unstable visual conditions, potentially generating false positive predictions (Baur et al., 2025). To improve system reliability, an additional validation mechanism is required. Previous studies reported that temporal validation methods in video-based detection systems can improve prediction consistency while reducing fluctuations in dynamic environments [9], [10].

Unlike fully autonomous surveillance systems, this study adopts a human-in-the-loop approach where detection results that have passed temporal validation are first displayed to the operator before any further response is executed. The audio warning module mounted on the drone is not activated automatically, but instead triggered manually by the operator using a control button. This mechanism is intended to maintain a balance between AI automation and human supervision, thereby minimizing the possibility of incorrect activation during surveillance operations.

In addition, ESP-NOW wireless communication is utilized to transmit trigger commands from the image processing system to the embedded module installed on the drone. ESP-NOW is considered suitable for real-time applications because it offers lightweight communication with low transmission latency [11], [12]. Through this integration, the proposed system not only performs object detection but also functions as a UAV-based decision support system equipped with a controlled response mechanism.

A comparison between previous studies and the proposed research based on several key aspects, including detection method, UAV platform, wireless communication technology, temporal validation, human-in-the-loop capability, and overall contribution. Previous studies by Saeed et al. [7] and Tang et al. [2] primarily focused on UAV-based human detection using deep learning approaches,

namely YOLO and CNN, respectively. Although both studies demonstrated the effectiveness of aerial human detection, neither incorporated wireless communication mechanisms, temporal validation, nor human-in-the-loop functionality. Meanwhile, the studies conducted by Hailan et al. [11] and Wedyanti et al. [12] emphasized the implementation of ESP-NOW as a low-latency wireless communication protocol for embedded systems, but they did not integrate computer vision algorithms or UAV platforms for surveillance applications. In contrast, the proposed research combines the strengths of these previous works by integrating the YOLOv11 object detection model with a DJI FPV UAV platform and ESP-NOW wireless communication to enable real-time aerial surveillance. Furthermore, unlike previous studies, the proposed system incorporates temporal validation to improve detection reliability over consecutive frames and implements a human-in-the-loop mechanism that allows operators to verify and respond to detected targets before further actions are taken. Consequently, this research contributes a more comprehensive and practical UAV surveillance framework by integrating intelligent detection, efficient wireless communication, temporal consistency analysis, and human-assisted decision-making into a single operational system. Summarizes the differences between previous studies and the proposed research. Previous studies mainly focused on either object detection techniques or wireless communication systems separately. Saeed et al. [7] and Tang et al. [2] investigated UAV-based human detection and aerial image analysis without integrating communication subsystems or human-centered decision-making mechanisms. Meanwhile, Hailan et al. [11] and Wedyanti et al. [12] concentrated on ESP-NOW communication performance for embedded systems but did not incorporate computer vision or UAV surveillance applications. Therefore, a research gap still exists regarding the integration of real-time human detection, decision support mechanisms, and low-latency wireless communication within a single UAV surveillance framework.

The novelty of this research lies in the integration of four main components, namely YOLOv11-based real-time human detection, temporal validation, human-in-the-loop confirmation, and ESP-NOW wireless communication within a single UAV surveillance framework. Unlike previous studies that focused primarily on object detection or wireless communication independently, the proposed system combines artificial intelligence, operator supervision, and low-latency communication to improve detection reliability while maintaining human control over response activation. From a conceptual perspective, the proposed system represents a cyber-physical architecture consisting of sensing components (UAV camera), processing units (YOLOv11), temporal validation mechanisms, wireless communication modules (ESP-NOW), and actuation systems (audio warning module), while preserving human operators as the final decision-makers [13]. Therefore, this study aims to develop a real-time human detection system using YOLOv11 integrated with human-in-the-loop confirmation and ESP-NOW communication for FPV drone surveillance applications.

2. Research Methodology

System Architecture

The proposed system architecture consisted of several main components, namely the DJI FPV drone, YOLOv11 detection model, laptop-based processing system, ESP32 modules, DFPlayer Mini audio module, and ESP-NOW wireless communication. The DJI FPV drone functioned as the primary aerial image acquisition device that transmitted live video feeds during surveillance operations. The video stream was processed in real time on a laptop using Visual Studio Code (VSCode), Open Broadcaster Software (OBS) Studio, and virtual camera configuration through Zadig Driver.

The YOLOv11 model was responsible for detecting human objects from aerial imagery in real time. Object detection is a fundamental task in computer vision that identifies object classes and determines their locations through bounding boxes simultaneously within images or video streams [4]. The YOLO family is widely recognized for achieving high inference speed while maintaining detection accuracy, making it suitable for UAV surveillance systems operating in dynamic environments [5], [14]. In addition, YOLOv11 provides improvements in computational efficiency and small-object detection capability for aerial imagery applications [15], [16].

To improve detection stability, the system implemented a temporal validation mechanism before generating notifications for the operator. Transient detections in real-time video systems may produce false positive predictions caused by rapid movement, pose variation, and dynamic visual conditions [8]. Previous studies also reported that temporal validation approaches can improve prediction consistency and reduce fluctuations in video-based detection systems [9], [10].

Unlike fully autonomous surveillance systems, this research adopted a human-in-the-loop approach where validated detection results were first displayed to the operator before any response mechanism was activated. Human-in-the-loop systems combine artificial intelligence automation with human decision-making to improve reliability and minimize incorrect activation during security surveillance operations [17].

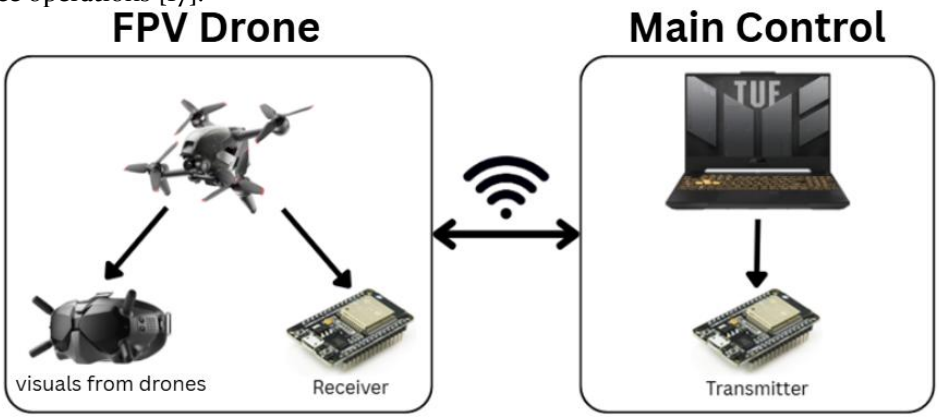

Figure 1. Proposed System Architecture

Figure 1 illustrates the overall architecture of the proposed cyber-physical surveillance system consisting of sensing, processing, communication, and actuation layers integrated with human operator control.

Dataset Acquisition

The dataset used in this study consisted of primary and secondary data. Primary data were collected directly using a DJI FPV Combo drone during outdoor surveillance simulations under real operational conditions. The captured aerial imagery contained variations in altitude, object scale, camera movement, and environmental lighting to represent dynamic FPV surveillance conditions. Secondary data were obtained from publicly available datasets on Kaggle and GitHub related to aerial human detection.

The use of UAV aerial imagery introduces several technical challenges. Objects captured from top-view drone perspectives often appear very small, causing weak feature representation during neural network processing [18]. In addition, FPV drones inherently produce motion blur and dynamic perspective changes due to rapid movement and camera ego-motion [2]. These characteristics increase the difficulty of stable real-time human detection.

This study applied a multi-class training strategy although the final inference system only displayed a single-class output, namely humans. Multi-class training enabled the model to learn visual differences between humans and surrounding objects such as vehicles, roads, and buildings, thereby reducing false positive predictions during inference.

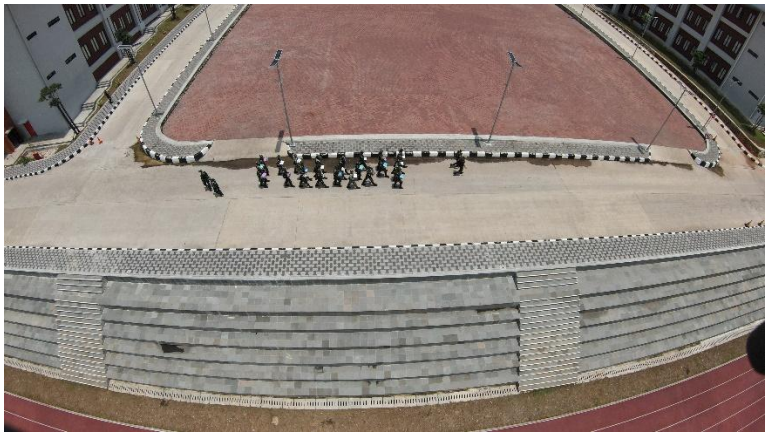

Figure 2. DJI FPV Drone Used for Dataset Collection

Figure 2 shows the DJI FPV Combo drone used for collecting primary aerial imagery datasets during surveillance simulations. The drone provided dynamic FPV video streams suitable for real-time human detection experiments.

Table 2. Dataset Composition

No.	Dataset Category	Number of Images	Percentage (%)
1.	Training Set	700	70%
2.	Validation Set	200	20%
3.	Testing Set	100	10%
4	Total	1000	100%

The dataset consisted of 1000 aerial images containing annotated human objects collected from both primary and secondary sources. The dataset was divided into training, validation, and testing subsets with proportions of 70%, 20%, and 10%, respectively. The training set contained 700 images and was used to optimize the YOLOv11 model parameters. Meanwhile, the validation and testing sets consisted of 200 and 100 images, respectively, and were utilized to monitor model convergence and evaluate the generalization capability of the trained model. The combination of primary and secondary datasets increased image diversity and improved the robustness of the proposed human detection system under various environmental conditions.

Table 3. Dataset Source

No.	Dataset Type	Source	Description
1.	Primary Dataset	DJI FPV Drone	Aerial images captured during surveillance simulations
2.	Secondary Dataset	Kaggle	Public aerial human detection dataset
3.	Secondary Dataset	GitHub	Public UAV human detection dataset

Primary data were collected using a DJI FPV Combo drone during outdoor surveillance simulations. The captured images contained variations in altitude, lighting conditions, object scale, and viewing angles to represent dynamic FPV surveillance scenarios. In addition, secondary datasets were obtained from publicly available sources, including Kaggle and GitHub repositories, to increase dataset diversity and improve the generalization capability of the YOLOv11 model.

Data Annotation and Augmentation

All datasets were annotated using the Roboflow platform with YOLO annotation format. Human objects were manually labeled using bounding boxes to generate accurate ground-truth data for supervised learning. Annotation quality plays an important role in object detection systems because

inaccurate bounding boxes or labeling inconsistencies may reduce model performance and increase classification errors [19].

Roboflow was selected because it provides integrated tools for annotation, dataset management, augmentation, and format conversion compatible with YOLO-based object detection systems. Automated augmentation platforms such as Roboflow have been shown to improve dataset diversity and detection performance in deep learning applications [20].

To improve generalization capability and reduce overfitting, several augmentation techniques were applied, including rotation, horizontal flipping, brightness adjustment, contrast enhancement, and scaling transformation. Data augmentation is important for aerial imagery because drone datasets often contain limited environmental variations and dynamic perspectives [21]. In addition, advanced augmentation strategies such as Mosaic augmentation have been reported to improve small-object detection sensitivity in aerial datasets [22].

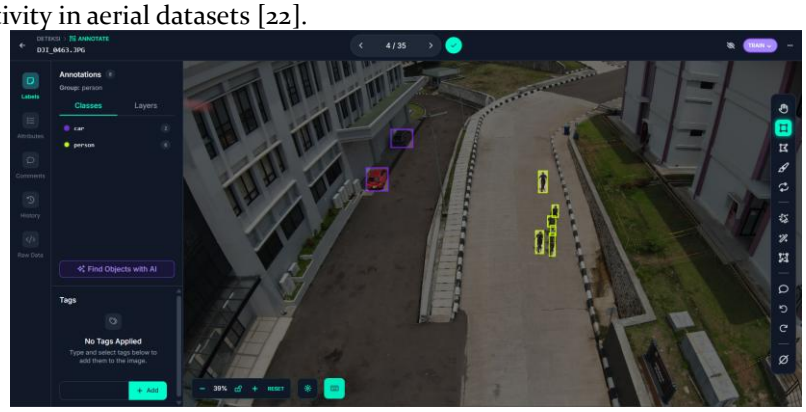

Figure 3. Annotation Process Using Roboflow

Figure 3 presents the annotation process conducted using Roboflow to generate YOLO-compatible ground-truth labels for human detection.

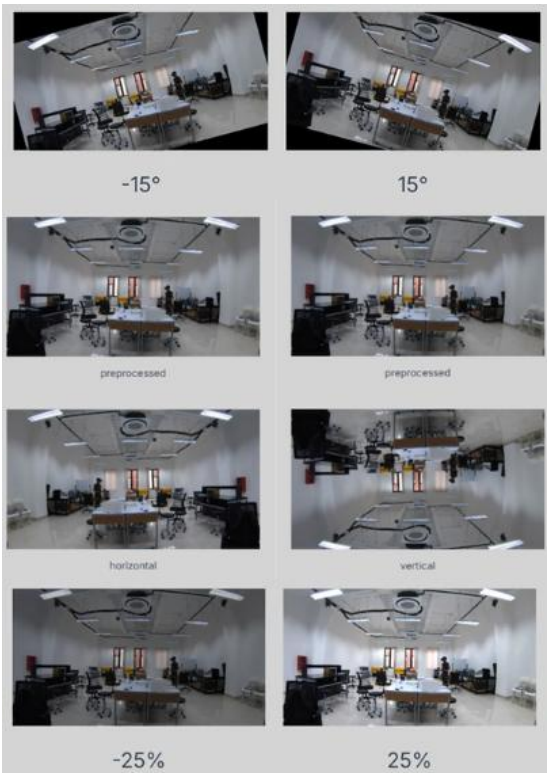

Figure 4. Data Augmentation Process

Figure 4 demonstrates several augmentation techniques implemented in this research, including rotation, flipping, brightness adjustment, contrast enhancement, and scaling transformation to improve dataset diversity.

YOLOv11 Training

The YOLOv11 model training process was conducted using Google Colab with GPU acceleration to improve computational efficiency. The training process involved configuring several hyperparameters such as epoch number, image size, batch size, and learning rate.

During training, model convergence was monitored using train loss and validation loss curves. The total loss value consisted of several components, namely box loss, classification loss, and Distribution Focal Loss (DFL). Box loss measures localization accuracy between predicted and ground-truth bounding boxes, while classification loss evaluates object class prediction performance [23], [24]. Distribution Focal Loss improves localization precision by modeling object boundary positions as discrete distributions [25], [26].

Model performance was evaluated using precision, recall, F1-score, and mean Average Precision (mAP). Precision measures the proportion of correctly predicted positive detections, whereas recall evaluates the ability of the model to detect all relevant objects [27], [28]. F1-score represents the harmonic balance between precision and recall (Yacouby et al., 2020). Meanwhile, mAP@0.5 and mAP@0.5:0.95 are widely used as standard evaluation metrics in modern object detection systems [29], [30].

Table 4. YOLOv11 Training Configuration

No	Parameter	Value
1	Model	YOLOv11
2	Epoch	100
3	Batch Size	16
4	Image Size	640x640
5	Platform	Google Colab
6	Learning Rate	0.01
7	Optimizer	SGD
8	Scheduler	Cosine
9	Framework	Ultralytics YOLO
10	GPU	NVIDIA Tesla T4

The YOLOv11 model was trained using the Ultralytics framework in the Google Colab environment with GPU acceleration. The training process utilized an NVIDIA Tesla T4 GPU to accelerate model optimization and reduce training time. Several hyperparameters, including learning rate, optimizer, scheduler, batch size, and image resolution, were configured to achieve stable convergence and improve detection performance.

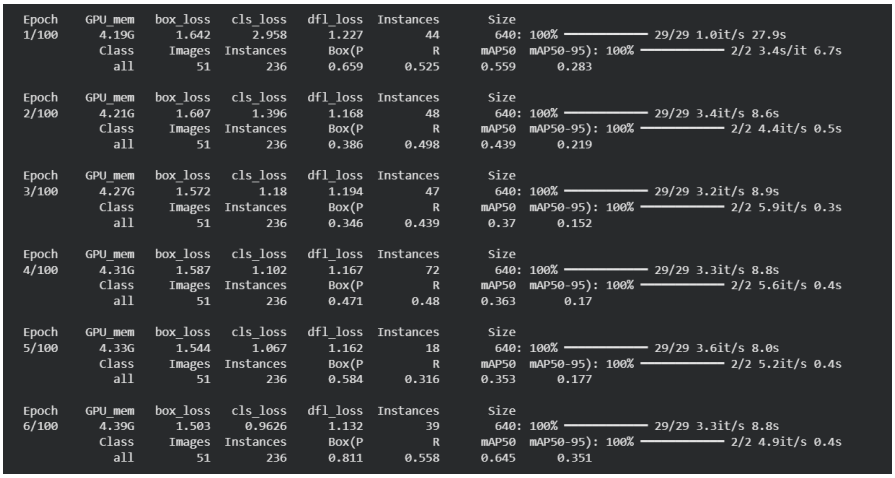

Figure 5. YOLOv11 Training Process in Google Colab

Figure 5 illustrates the YOLOv11 training process performed in Google Colab using GPU acceleration.

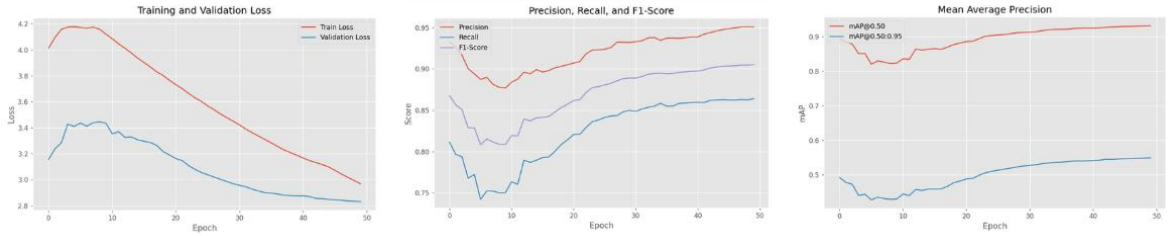

Figure 6. Training and Validation Performance Curves

Figure 6 presents the training and validation performance of the YOLOv11 model, including loss values, precision, recall, F1-score, and mean Average Precision (mAP). The training and validation loss curves exhibit a consistent downward trend throughout the training process, indicating stable model convergence without significant overfitting. At the end of training, the training loss decreased to approximately 3.0, while the validation loss converged to around 2.8.

The precision, recall, and F1-score curves initially decreased during the early epochs before gradually improving and stabilizing as the model learned more representative features from the aerial dataset. The final precision value reached approximately 0.95, while recall and F1-score converged to around 0.86 and 0.90, respectively. These results indicate that the proposed model achieved a balanced performance between minimizing false positive detections and maximizing object detection capability.

In addition, the mAP@0.5 and mAP@0.5:0.95 curves showed progressive improvements during training. The final mAP@0.5 value reached approximately 0.92, whereas mAP@0.5:0.95 converged to around 0.55. The higher mAP@0.5 value demonstrates that the model was able to localize human objects accurately at a standard Intersection over Union (IoU) threshold, while the lower mAP@0.5:0.95 reflects the increasing difficulty of achieving accurate localization under stricter IoU requirements. Overall, the results indicate that YOLOv11 achieved stable convergence and demonstrated good generalization capability for real-time aerial human detection applications.

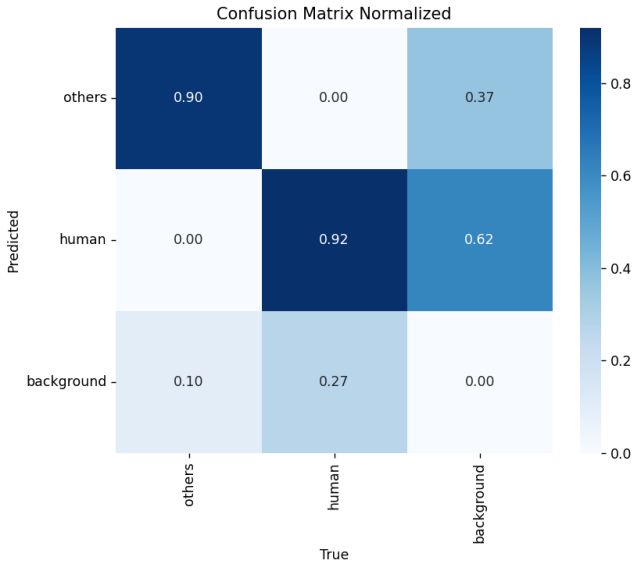

Figure 7. Confusion Matrix of YOLOv11 Model

Figure 7 presents the normalized confusion matrix of the YOLOv11 model after the training process. The confusion matrix was used to evaluate the classification capability of the model by comparing predicted labels with the corresponding ground-truth labels. The results indicate that the YOLOv11 model achieved high classification performance for both the human and others classes. Specifically, the model correctly classified approximately 92% of human objects and 90% of objects belonging to the others class. These results demonstrate that the model was able to learn discriminative visual features effectively and distinguish human objects from other surrounding objects in aerial imagery.

However, a small proportion of human objects was still classified as background, which may be caused by the limited size of objects in aerial images, variations in viewing angles, and environmental lighting conditions. Similar challenges have been reported in previous UAV-based object detection studies, where small-object representation remains one of the main limitations in aerial surveillance systems. Nevertheless, the overall confusion matrix demonstrates that the proposed YOLOv11 model provides robust classification performance and is suitable for real-time human detection applications using FPV drones.

Real-Time Detection System

The real-time detection system was implemented using Visual Studio Code integrated with OBS Studio and Zadig Driver. The FPV live video feed transmitted from the DJI FPV drone was connected to the processing system through virtual camera configuration for continuous inference. The YOLOv11 model continuously processed incoming video frames to detect human objects in real time. CNN-based deep learning architectures are highly effective for processing aerial imagery because they can automatically learn hierarchical visual features from complex image data (Hua et al., 2025; Deng et al., 2025). In UAV surveillance systems, CNN models have demonstrated strong capability in detecting small objects and handling dynamic environmental conditions [31], [32]. To minimize unstable detections, temporal validation was implemented before generating notifications. This mechanism evaluated detection consistency across multiple consecutive frames to reduce transient detections and false positive activations during real-time operation.

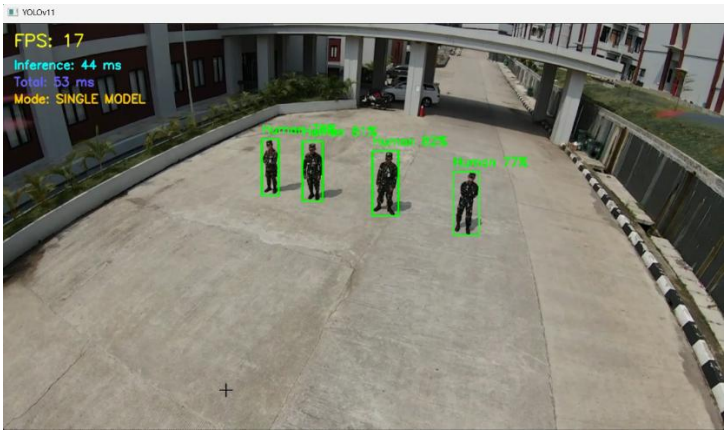

Figure 8. Real-Time Human Detection System

Figure 8 illustrates the implementation of the proposed real-time detection system during FPV drone surveillance operations.

ESP-NOW Communication System

ESP-NOW communication was implemented to support lightweight and low-latency wireless communication between the processing system and the embedded module mounted on the drone. ESP-NOW is a connectionless peer-to-peer communication protocol developed by Espressif Systems that enables direct data transmission between ESP devices without requiring external network infrastructure [11]. Compared to conventional Wi-Fi communication, ESP-NOW eliminates network association and Internet Protocol configuration processes, thereby reducing communication overhead and improving transmission latency [12]. The protocol also provides low power consumption and stable short-range communication, making it suitable for UAV-based embedded systems. Two ESP32 modules were configured as transmitter and receiver nodes. The transmitter ESP32 received trigger commands from the laptop-based detection system after operator confirmation was performed. The signal was then transmitted wirelessly to the receiver ESP32 mounted on the drone. After receiving the signal, the receiver activated the DFPlayer Mini module connected to a speaker to generate an audio warning response.

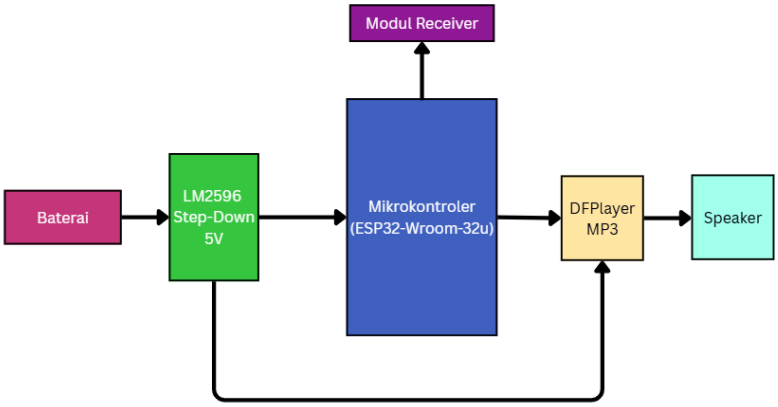

Figure 9. ESP-NOW Communication Architecture

Figure 9 presents the ESP-NOW communication architecture implemented in the proposed surveillance system.

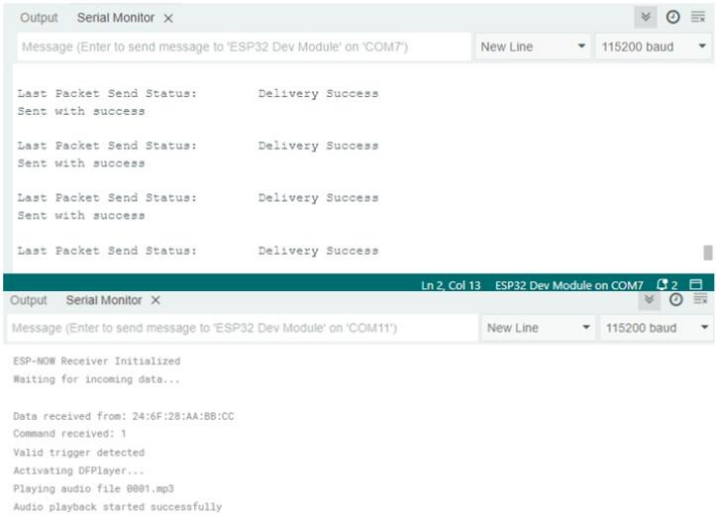

Figure 10. ESP32 Signal Transmission Process

Figure 10 illustrates the process of signal transmission and reception between ESP32 modules during real-time operation.

Table 5. ESP32 Receiver Pin Configuration

No	Component	Component Pin	Connected to ESP32 Pin
1	DFPlayer Mini	TX	GPIO16 (RX2)
2	DFPlayer Mini	RX	GPIO17 (TX2)
3	DFPlayer Mini	VCC	5V Output of LM2596
4	DFPlayer Mini	GND	GND ESP32
5	Speaker	SPK_1	DFPLAYER Output (+)
6	Speaker	SPK_2	DFPLAYER Output (-)
7	LM2596	OUT (+)	5V Pin of ESP32
8	LM2596	OUT (-)	GND ESP32
9	Li-ion Battery	(+)	IN (+) of LM2596
10	Li-ion Battery	(-)	IN (-) of LM2596

System Evaluation

The proposed system was evaluated under several environmental conditions, including morning, afternoon, and evening scenarios. The evaluation focused on analyzing the capability of YOLOv11 in detecting human objects under different lighting conditions, object distances, and movement dynamics.

The evaluation process analyzed precision, recall, F1-score, and mAP metrics, as well as the effectiveness of temporal validation and human-in-the-loop mechanisms in minimizing false positive activation during real-time implementation.

Table 6. Human Detection Testing Scenarios

No	Scenario	Time Condition	Detection Target
1	Test 1	Morning	Human
2	Test 2	Afternoon	Human
3	Test 3	Evening	Human

The testing scenarios were designed to evaluate system robustness and detection stability under environmental conditions commonly encountered during FPV drone surveillance operations.

3. Result and Discussion

Real-Time Human Detection Performance

The performance of the proposed system was evaluated through real-time human detection experiments using a DJI FPV drone under different environmental conditions. Testing was conducted during morning, afternoon, and evening periods to analyze the robustness of the YOLOv11 model under varying illumination levels and outdoor conditions.

Table 7. Human Detection Performance Under Different Environmental Conditions

No	Environmental Condition	Average Confidence (%)	Detection Success Rate (%)
1	Morning	81,25%	71,13%
2	Afternoon	78,38%	75,94%
3	Evening	79,88%	78,03%

Figure 11. Human Detection Result During Morning Condition

Figure 12. Human Detection Result During Afternoon Condition

Figure 13. Human Detection Result During Evening Condition

Based on the experimental results presented in Table 7, the proposed YOLOv11-based detection system successfully identified human objects under all testing conditions. The system maintained stable detection performance despite variations in environmental illumination and object appearance. During morning testing, the detection model achieved stable performance due to sufficient natural lighting and minimal shadow interference. Human objects were detected consistently, indicating that the trained model was capable of extracting relevant visual features from aerial imagery. The afternoon scenario generally produced the highest confidence values. This condition can be attributed to optimal illumination intensity, which enhanced object visibility and improved feature representation within the input frames. Consequently, the model generated more confident predictions compared to other testing periods. In the evening scenario, confidence values slightly decreased because of reduced lighting intensity and increasing shadow effects. Nevertheless, the detection system remained capable of identifying human objects accurately. These results demonstrate that YOLOv11 possesses strong robustness against moderate environmental lighting variations, making it suitable for UAV-based surveillance applications. The findings are consistent with previous studies indicating that modern YOLO architectures provide effective object detection performance while maintaining real-time processing capability in dynamic aerial environments [5], [16].

Single and Multi Detection Analysis

To further evaluate the detection capability of the proposed system, experiments were conducted using both single and multi-detection scenarios. The objective was to analyze the model's ability to detect varying numbers of human targets within the same frame.

Table 8. Detection Results for Different Numbers of Persons

No	Number of Humans	Detection Status
1	1	Successful
2	2	Successful
3	3	Successful
4	4	Successful

Figure 14. Single-Detection Result

Figure 15. Multi-Detection Result

The experimental results indicate that the proposed YOLOv11 model successfully detected human objects across different crowd densities. In the single-detection scenario, the model consistently produced high-confidence predictions because the target occupied a larger portion of the image and experienced minimal visual interference. In multi-detection scenarios, the model was able to identify multiple individuals simultaneously within the same frame. However, variations in confidence values were observed depending on object scale, distance from the drone, overlap between individuals, and viewing angle. Objects located farther from the camera generally produced lower confidence values because they occupied fewer pixels within the image. Despite these challenges, the model maintained reliable detection performance, indicating effective feature extraction and object localization capability. These findings demonstrate that the proposed system is capable of supporting practical surveillance operations where the number of observed individuals may vary dynamically.

ESP-NOW Communication Performance

The communication subsystem was evaluated to determine the reliability of ESP-NOW for transmitting activation commands between the processing system and the embedded response module mounted on the drone.

Table 9. ESP-NOW Communication Test Results

No	Distance (m)	Communication Status	Delay
1	0 m	Successful	7 ms
2	100 m	Successful	12 ms
3	200 m	Successful	24 ms
4	300 m	Successful	53 ms

Figure 16. ESP-NOW Communication Testing Scenario

The results presented in Table 5 demonstrate that ESP-NOW successfully transmitted activation signals throughout all testing scenarios. Communication remained stable without requiring external network infrastructure such as routers or internet connectivity. The low-latency characteristics of ESP-NOW contributed significantly to the responsiveness of the proposed system. Following operator confirmation, activation commands were delivered rapidly to the receiver module, allowing the DFPlayer Mini to generate audio responses in near real-time. The communication protocol also exhibited stable performance during outdoor testing, making it suitable for integration with UAV-based surveillance systems. These findings support previous studies that identified ESP-NOW as an effective communication protocol for low-power embedded applications requiring fast and reliable wireless transmission [11], [12].

Temporal Validation and Human-in-the-Loop Analysis

One of the primary contributions of this research is the integration of temporal validation and human-in-the-loop mechanisms within the UAV surveillance framework. In conventional real-time object detection systems, temporary visual disturbances, motion blur, and rapid object movements may generate false positive predictions. To address this issue, the proposed system employed temporal validation, where detected objects were required to remain consistently visible across multiple consecutive frames before generating notifications. This mechanism effectively filtered transient detections and improved prediction stability. By requiring temporal consistency, the system reduced the probability of triggering responses based on momentary or unreliable detections. After temporal validation was satisfied, the system did not automatically activate the audio response module. Instead, the validated detection result was presented to the operator for confirmation. This human-in-the-loop mechanism ensured that the final decision remained under human supervision. The implementation of human-in-the-loop control offers several advantages. First, it minimizes the risk of false activation caused by incorrect detections. Second, it allows operators to evaluate contextual information that may not be fully captured by the artificial intelligence model. Third, it maintains accountability and operational control during surveillance activities. Consequently, the proposed system functions as an AI-assisted surveillance platform rather than a fully autonomous surveillance system. This design aligns with current trends in intelligent security systems, where artificial intelligence is used to support human decision-making rather than replace it entirely [17].

Discussion

The experimental results demonstrate that the integration of YOLOv11, temporal validation, human-in-the-loop confirmation, and ESP-NOW communication can effectively support real-time UAV surveillance operations. The YOLOv11 model successfully detected human objects under varying environmental conditions while maintaining stable performance during real-time operation. The ability to detect both single and multiple individuals indicates that the model possesses sufficient robustness for practical surveillance scenarios. Furthermore, the implementation of ESP-NOW communication enabled reliable wireless transmission of activation signals with minimal delay. This capability is important for surveillance applications requiring immediate responses following object detection. Compared to previous UAV-based detection systems that primarily focus on object recognition accuracy, the proposed system introduces an additional decision-support layer through temporal validation and operator confirmation. Overall, the proposed framework demonstrates the feasibility of combining deep learning-based object detection, low-latency wireless communication, and human-centered decision-making into a single integrated UAV surveillance system. The results suggest that the system has strong potential for deployment in campus security monitoring, perimeter patrol operations, and other intelligent aerial surveillance applications. Nevertheless, several limitations remain in the proposed system. Detection performance may decrease at higher drone altitudes because human objects occupy fewer pixels in aerial images. In addition, extreme weather conditions such as rain, fog, and strong winds were not evaluated in this study. Future research should investigate more robust detection techniques and adaptive sensing strategies to improve system performance under challenging environmental conditions.

4. Conclusion

This study developed and implemented a real-time human detection system for FPV drone surveillance by integrating the YOLOv11 object detection model with ESP-NOW wireless communication. The proposed system successfully detected human objects from aerial imagery captured by a DJI FPV drone under different lighting conditions, including morning, afternoon, and evening scenarios. Experimental results demonstrated that YOLOv11 achieved reliable real-time detection performance while maintaining stable operation throughout surveillance activities. The incorporation of temporal validation enhanced detection reliability by filtering transient detections and reducing false positives, whereas the human-in-the-loop mechanism ensured that final activation decisions remained under operator supervision, allowing artificial intelligence to function as a decision-support tool rather than replacing human judgment. In addition, the ESP-NOW communication subsystem enabled low-latency and stable transmission of activation signals between the processing unit and the embedded response module. The primary contribution of this research is the integration of YOLOv11-based human detection, temporal validation, human-in-the-loop confirmation, and ESP-NOW communication into a unified UAV surveillance framework. Unlike previous studies that addressed either object detection or wireless communication independently, this work combines intelligent detection, operator-centered decision-making, and efficient communication to improve surveillance reliability and responsiveness. Despite these contributions, several limitations remain. The system was evaluated only for human detection under limited environmental conditions and still depends on an external processing unit. Future research should extend the framework to support multi-object detection, integrate edge computing directly on UAV platforms, incorporate object tracking algorithms, and investigate thermal imaging for improved performance under low-light and nighttime conditions. These enhancements would further increase the robustness, mobility, and practical applicability of intelligent UAV surveillance systems across a wider range of operational environments.

References

[1] A. Polina, H. Suparwito, and R. Kumalasanti, "Aerial object detection analysis: Challenges and preliminary results," *E3S Web Conf.*, vol. 475, Jan. 2024, doi: 10.1051/e3sconf/202447502017.

[2] G. Tang, J. Ni, Y. Zhao, Y. Gu, and W. Cao, "A survey of object detection for UAVs based on deep learning," *Remote Sens.*, vol. 16, no. 1, p. 149, 2023.

[3] J. Zhang, G. Wan, M. Jiang, G. Lu, X. Tao, and Z. Huang, "Small object detection in UAV image based on improved YOLOv5," *Syst. Sci. Control Eng.*, vol. 11, no. 1, p. 2247082, 2023.

[4] L. Jiao and M. I. Abdullah, "YOLO series algorithms in object detection of unmanned aerial vehicles: a survey," *Serv. Oriented Comput. Appl.*, vol. 18, no. 3, pp. 269–298, 2024.

[5] S.-Y. Yang, H.-Y. Cheng, and C.-C. Yu, "Real-time object detection and tracking for unmanned aerial vehicles based on convolutional neural networks," *Electronics*, vol. 12, no. 24, p. 4928, 2023.

[6] Z. Xu, H. Zhao, P. Liu, L. Wang, G. Zhang, and Y. Chai, "SRTSOD-YOLO: stronger real-time small object detection algorithm based on improved YOLO11 for UAV imageries," *Remote Sens.*, vol. 17, no. 20, p. 3414, 2025.

[7] Z. Saeed, M. H. Yousaf, R. Ahmed, S. A. Velastin, and S. Viriri, "On-board small-scale object detection for unmanned aerial vehicles (UAVs)," *Drones*, vol. 7, no. 5, p. 310, 2023.

[8] J. Baur and F. O. Nitsche, "A False-Positive-Centric Framework for Object Detection Disambiguation," *Remote Sens.*, vol. 17, no. 14, p. 2429, 2025.

[9] S. K. Dubey, J. V Satyanarayana, and C. K. Mohan, "False positive elimination in object detection methods for videos," in *Sixteenth International Conference on Machine Vision (ICMV 2023)*, SPIE, 2024, pp. 146–153.

[10] F. Lu, C. Zeng, H. Shi, Y. Xu, and S. Fu, "Real-Time Detection Sensor for Unmanned Aerial Vehicle Using an Improved YOLOv8s Algorithm," *Sensors*, vol. 25, no. 19, p. 6246, 2025.

[11] M. A. Hailan, N. M. Ghazaly, and B. M. Albaker, "ESPNow Protocol-Based IIoT System for Remotely Monitoring and Controlling Industrial Systems," *J. Robot. Control*, vol. 5, no. 6, pp. 1924–1942, 2024.

[12] F. P. Wedyanti and A. Wagyana, "Analisis Performasi ESP-NOW Sebagai Protokol Komunikasi Efisien Antar Multi-node dan Gateway dalam Sistem IoT," in *Seminar Nasional Inovasi Vokasi*, 2025, pp. 607–614.

[13] H. Herdianto, H. Hafni, D. Nasution, and S. Ramadhan, "Implementasi metode YOLO pada deteksi objek manusia," *METHOMIKA J. Manaj. Inform. Komputerisasi Akunt.*, vol. 8, no. 2, pp. 234–240, 2024.

[14] P. Sharma, R. Tyagi, and P. Dubey, "Optimizing Real-Time Object Detection-A Comparison of YOLO Models.," *Int. J. Innov. Res. Comput. Sci. Technol.*, vol. 12, no. 3, pp. 57–74, 2024.

[15] R. Khanam, T. Asghar, and M. Hussain, "Comparative performance evaluation of yolov5, yolov8, and yolov11 for solar panel defect detection," in *Solar*, MDPI, 2025, p. 6.

[16] R. Sapkota, Z. Meng, M. Churuvija, X. Du, Z. Ma, and M. Karkee, "Comprehensive performance evaluation of yolov12, yolo11, yolov10, yolov9 and yolov8 on detecting and counting fruitlet in complex orchard environments," *arXiv Prepr. arXiv2407.12040*, 2024.

[17] E. Mosqueira-Rey, E. Hernández-Pereira, D. Alonso-Ríos, J. Bobes-Bascarán, and Á. Fernández-Leal, "Human-in-the-loop machine learning: a state of the art," *Artif. Intell. Rev.*, vol. 56, no. 4, pp. 3005–3054, 2023.

[18] L. Zhu, J. Xiong, F. Xiong, H. Hu, and Z. Jiang, "Yolo-drone: Airborne real-time detection of dense small objects from high-altitude perspective," *arXiv Prepr. arXiv2304.06925*, 2023.

[19] J. Nassar, V. Pavon-Harr, M. Bosch, and I. McCulloh, "Assessing data quality of annotations with Krippendorff alpha for applications in computer vision," *arXiv Prepr. arXiv1912.10107*, 2019.

[20] U. Nisa, "Image augmentation approaches for small and tiny object detection in aerial images: A review," *Multimed. Tools Appl.*, vol. 84, no. 19, pp. 21521–21568, 2025.

[21] X. Hao, L. Liu, R. Yang, L. Yin, L. Zhang, and X. Li, "A review of data augmentation methods of remote sensing image target recognition," *Remote Sens.*, vol. 15, no. 3, p. 827, 2023.

[22] A. M. Abdulghani, M. M. Abdulghani, W. L. Walters, and K. H. Abed, "Multiple Data Augmentation Strategy for Enhancing the Performance of YOLOv7 Object Detection Algorithm.," *J. Artif. Intell.*, vol. 5, 2023.

[23] D. Cai, Z. Zhang, and Z. Zhang, "Corner-point and foreground-area IoU loss: Better localization of small objects in bounding box regression," *Sensors*, vol. 23, no. 10, p. 4961, 2023.

[24] X. Qian, S. Gao, W. Deng, and W. Wang, "Improving oriented object detection by scene classification and task-aligned focal loss," *Mathematics*, vol. 12, no. 9, p. 1343, 2024.

[25] Y. Ding *et al.*, "Application of Improved YOLOv8 Image Model in Urban Manhole Cover Defect Management and Detection: Case Study," *Sensors*, vol. 25, no. 13, p. 4144, 2025.

[26] M. Liu, C. Zhang, and C. Lin, "GAB-YOLO: a lightweight deep learning model for real-time detection of

abnormal behaviors in juvenile greater amberjack fish," *Front. Mar. Sci.*, vol. 12, p. 1574580, 2025.
[27] A. A. Adegun, J. V. Fonou Dombeu, S. Viriri, and J. Odindi, "State-of-the-art deep learning methods for objects detection in remote sensing satellite images," *Sensors*, vol. 23, no. 13, p. 5849, 2023.
[28] L. He, Y. Zhou, L. Liu, W. Cao, and J. Ma, "Research on object detection and recognition in remote sensing images based on YOLOv11," *Sci. Rep.*, vol. 15, no. 1, p. 14032, 2025.
[29] A. Rahman, Y. Lu, and H. Wang, "Performance evaluation of deep learning object detectors for weed detection for cotton," *Smart Agric. Technol.*, vol. 3, p. 100126, 2023.
[30] Y. Ma *et al.*, "I-YOLOv11n: A Lightweight and Efficient Small Target Detection Framework for UAV Aerial Images," *Sensors*, vol. 25, no. 15, p. 4857, 2025.
[31] L. Xu, Y. Zhao, Y. Zhai, L. Huang, and C. Ruan, "Small object detection in UAV images based on Yolov8n," *Int. J. Comput. Intell. Syst.*, vol. 17, no. 1, p. 223, 2024.
[32] 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.*, vol. 16, no. 22, p. 4265, 2024.