International Journal of Applied Mathematics

Volume 38 No. 9s , 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

AN ENHANCED PERSON DETECTION IN AERIAL SURVEILLANCE USING DEEP LEARNING TECHNIQUES

Renas Rajab Asaad [1,2], Prof. Dr. Subhi Rafeeq Mohammed [3]

[1] Department of Information Technology, Technical College of Duhok, Duhok Polytechnic University

[2] Department of Information Technology, Technical College of Informatics - Akre, Akre University for Applied Sciences

[3] Cybersecurity Engineering Department, Technical College of Engineering, Duhok Polytechnic University

Email: renas.rajab@auas.edu.krd

Abstract:

Person detection in aerial surveillance images is quite special due to the small size of objects, complex backgrounds, and changeable scales. Deep learning is a strong solution for object detection using Convolutional Neural Networks (CNNs) as the backbone. Beyond that, there are the several architectures for CNNs such as Visual Geometry Group (VGG), Residual Network (ResNet), Densely Connected Convolutional Networks (DenseNet) which turn out to have prominent performances in feature extraction and capturing spatial hierarchies. These models are fine-tuned for the specific characteristics of the aerial images. The better improvements of detection performance were observed in region-based CNNs such as Faster Region-based CNN (R-CNN) which uses Region Proposal Network(RPN) to achieve better efficiency in detection and accuracy, while Mask R-CNN adds a segmentation mask to identify the exact shape of detected persons. Single Shot Detectors (SSD) and You Only Look Once (YOLO) are fast and real-time. SSDs balance speed with accuracy, while YOLO detects objects by predicting bounding boxes over a grid in a single pass. Moreover, attention mechanisms and Vision Transformers boost person detection by making models focus on relevant image regions and capturing long-range dependencies, thus improving the detection performance in cluttered scenes. Models like YOLOv4, YOLOv5, and YOLOv7 are the most preferred ones over others in object detection due to the real-time performance with high speed. While others, like Faster R-CNN and Mask R-CNN, have several stages of processing that take a longer time to respond and YOLO has a one neural network that identifies the bounding boxes and class probabilities, thus attaining a high frames-per-second. This comes with high accuracy (94.0%) and low loss (0.22) as indicated by impressive performance metrics of YOLOv7.

Keywords: Aerial Surveillance, Person Detection, Deep Learning, Convolutional Neural Networks (CNNs), Feature Pyramid Networks (FPN), Vision Transformers (ViTs), Object Detection.

Received: August 09 2025

Volume 38 No. 9s , 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

I. INTRODUCTION

Aerial surveillance has emerged as the key technology for very diversified domains related to security, disaster management, and urban planning. The growing utilization of drones, satellites, and other aerial platforms in the past few years has created a subsequent increase in demand for automated systems that allow real-time person detection. Person detection from aerial images is intrinsically very challenging due to several factors. This is further complicated by small object size, cluttered and complex backgrounds, and changes in scale, angle, and lighting. The task will therefore need efficient and robust algorithms that very accurately detect a person despite the challenges. Traditional image processing techniques often prove inadequate for the purpose and underscore the need for advanced methodologies in order to get reliable detection [1].

Convolutional Neural Networks has models like CNNs, VGG, ResNet, and DenseNet, are used to extract complex features from images, so that improved detection accuracy could be attained even in difficult conditions. Some of these methods, such as R-CNN, including its variants, Faster R-CNN and Mask R-CNN, use region proposal networks and segmentation masks to locate objects with higher accuracy, thus increasing detection performance. However, SSD and YOLO are more favored due to the real-time detection capabilities, thus creating a tradeoff midway in speed versus accuracy. Further improvements to FPN and ViTs are focused on improved detection of small objects and better contextual understanding, respectively [2,3].

There are several major challenges that occur in the effectiveness of person detection models in aerial surveillance. The first is the small size of a human body in an aerial image, which probably leads to missed detections due to the fact that many models fail. Moreover, the noise level and interference are boosted by cluttered and dynamic backgrounds in aerial images, as in the cases of urban scenes or dense vegetation. This easily leads to confusion between person and background in the model with large false positives and misclassifications. Another challenge lies in the variation in scale due to camera distance and angle that might affect detection performance on different scales. Occlusions, where people are partially concealed by objects or other people, further complicate tasks related to detection and localization. Likewise, very important are lighting and weather conditions, which range from shadows and glare to thick weather that can eventually occlude people and has a direct impact on the accuracy of person detection. Besides, state-of-the-art models, like Vision Transformers and Mask R-CNN, could be computationally heavy and turn out to be inappropriate in real-time applications on resource-constrained platforms [3,4]. The following illustrative contributions are derivable from the proposed person detection review in aerial surveillance using deep learning.

- The comparative analysis of the models like CNN, Region-based CNNs, Single Shot Detectors, You Only Look Once, and Vision Transformers is performed to select the bet model.

Volume 38 No. 9s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

- This review provides in-depth performance analysis of these models with respect to accuracy, speed, and robustness against challenging factors such as small object size, cluttered backgrounds, scale variability, and occlusions.
- It identifies and discusses key challenges and limitations of current person detection models in aerial surveillance, including issues related to small object detection, background interference, changes in scale, occlusions, and changes in environmental parameters.
- This paper also includes a number of potential solutions to the existing challenges and future research directions for improving the accuracy, efficiency, and generalization across diverse scenarios.

The structure of the paper is planned as follows. Section II gives the literature review on various models proposed by various researchers. Section III describes the architecture of different models with its advantages and disadvantages. Section IV shows the results obtained from various models. The conclusion of the review is given in Section V.

II. A REVIEW ON DEEP LEARNING MODELS

Koyun, O. C., et al. (2022) present a new framework designed to improve small object detection in aerial imagery. This approach involves the embedding of a focus mechanism to enhance resolution and the visualization of small objects that are usually hard to identify due to the small size and high altitude from which they are viewed. It thus sharpens the focus on small objects with a complex background, aiming to enhance both the accuracy and reliability of detection considerably [5]. Ahmed, I., et al. (2021) proposed an adaptation of the YOLOv3 model, driven by Gaussian features and transfer learning. This methodology modifies the architecture of YOLOv3 so that it handles overhead views of human detection. It enhances Gaussian-based features and employs transfer learning to realize improved accuracy and performance in smart city scenarios where efficient and accurate human detection is most required [6]. Azmat et al. (2023) proposed an elliptical modeling combined system for recognition from aerial surveillance images. The approach models elliptical shapes to enhance the detection and classification of various activities performed by humans, enhancing the recognition and identification capability of actions from aerial views [7]. Akshatha et al. (2023) propose a specific dataset that can be used to evaluate and benchmark algorithms for small object detection [8].

Bahmanyar et al. (2023) propose improvement in person detection from above for disaster response scenarios using deep neural networks. They target the enhancement of detection competency in such extreme scenarios of natural disasters where the identification of a person is prompt, relevant, and cardinal to effective rescue and response operations [9]. Khan, M. A., et al. (2023) designed as a lightweight model for real-time crowd density estimation in video surveillance applications. The model is aimed at being efficient and fast. It would work with live surveillance systems that seek fast yet accurate analysis over crowd conditions [10]. Thakur et al. (2023) proposed a framework for pedestrian detection in crowded scenes. The framework is designed to offer better accuracy and reliability of pedestrian detection within very crowded scenes, considering major challenges like

Volume 38 No. 9s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

occlusions and changes in scale that can improve the results of surveillance [11]. Sambolek, S., & Ivasic-Kos, M. (2024) explored methodologies that contain person detection in UAV aerial images [12].

Nguyen, H. et al. (2024) developed framework to bridge the gap between aerial and ground views of person re-identification. From these two views, the integrated features are made by advanced deep learning methodologies, aligning the varied data from different perspectives to enhance re-identifying individuals within diversified surveillance environments [13]. Janakiramiah et al. (2023) employs several deep learning models in order to upsurge the object detection abilities from aerial images, thus enhancing the performance of overall surveillance tasks that include the precise identification and tracking of objects within the captured footage [14]. Vasudevan, I., & Nithva, N. S. (2023) proposed a multi-model convolutional neural network for automatic person detection under search and rescue operations. In this sense, it is able to combine as many CNN models for elevated robust detection of individuals meant for search and rescue application and enable efficient and expedited identification during critical time [15]. Kim et al. (2023) proposed a model for UAV-surveillance that adaptively adjusts frame rates and selects a stabilized model for object recognition. The simplified approach identified the required small frame rates of the surveillance video and more appropriate object recognition model to be used for a given context of surveillance, mitigating problems relating to video stabilization and model [16].

Zhang et al. (2023) developed a method for small person detection by drone-based RGB and thermal imagery. This approach fuses RGB and thermal data in a framework that carries out detection on small or obscured person instances through deep learning techniques, thereby improving the accuracy and reliability in detecting tiny individuals from drone surveillance footage [17]. Zhang, H. et al. (2023) proposed an improved YOLOv5 and it has tremendously complex backgrounds, distinguished appearance diversity, and great scale variability of objects in aerial imagery [18]. Sunanda et al. (2023) suggested a Deep Mask R-CNN for vehicle detection in aerial surveillance images. The methodology aims to improve object detection and recognition as that can improve accuracies in detection and reduce false positives for vehicle identification from aerial footage [19]. Akinsemoyin et al. (2023) uses deep learning to track safety and health activities on a construction site. This is used to identify bad practices through aerial images and consequently ensure compliance with safety protocols in connection with the monitoring of the site's safety during construction [20].

Aibibu et al. (2023) proposed a Gaussian Wasserstein network to improve the detection of minor objects in low-resolution infrared images for unmanned aerial vehicle application in urban road surveillance. The Gaussian Wasserstein distance computation is used to represent facial features and enhances the feature representation in a low-resolution infrared image of small objects for urban road scenes [21]. Zhang, X. et al. (2023) proposed a person detection algorithm based on instance segmentation using Maximum Mean Discrepancy (MMD) distance. This method improves person detection by incorporating both instance segmentation, an approach that brings accurate object localization, and MMD distance, which minimizes domain shifting between the set of labeled and unlabeled data and detection in

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

aerial imagery [22]. Hanzla, M., et al. (2024) deployed U-NET segmentation in combination with DeepSORT for vehicle surveillance in aerial imagery. This approach relies on U-NET to attain the vehicle segmentation's accuracy and DeepSORT for robust tracking along frames to boost the accuracy of vehicle detection [23].

Stan, A. S., et al. (2023) addresses person detection and tracking using UAVs. The power of neural networks is tapped in the development of complex neural network models to raise the accuracy and efficiency in tracking people observed on UAV-recorded video by applying it in aerial surveillance applications [24]. Valarmathi, B. et al. (2023) proposed human detection and action recognition models during rescue operations through disasters. They implemented YOLOv3 architecture to detect humans and recognize the actions from aerial images, with the aim of making the response to the identification more effective during an emergency [25]. Zhang, Y. et al. (2023) enhance target detection algorithm on aerial images. Therefore, improved target detection was enhanced through the use of refined algorithm techniques designed to solve the intrinsic complicated challenges of varying sea conditions and dissipated targets that define a maritime environment [26]. Patel, N. et al (2023) designed framework of utilization using convolutional neural networks and unmanned aerial vehicles for the protection of human life. It is used to integrate CNN with UAV technology to monitor and respond to threats to public safety [27].

Kucukayan, G., & Karacan, H. (2024) present YOLO-IHD for indoor drones to the audience. This method enhances YOLO's in real-time human detection within indoor premises with the aim to tune the model appropriately for speed and precision under confined space conditions [28]. Abood et al. (2023) developed a border surveillance system using autonomous UAVs. This design is proposed with a vision of providing high surveillance by increasing the efficiency of border security operations using advanced autonomous systems that can monitor a wider area without compromise the type of data [29]. Xu, S. et al. (2024) searches for multi-grain focused mechanisms for person identification in aerial imagery. A multi-granularity-attentive network-based method for the task is proposed. This will enable the ability to revisit detailed and contextual information at the corresponding scales [30].

Rahman et al. (2023) proposed YOLOv8 with Wireless Sensor Networks (WSN) and UAV aerial photography. This approach takes advantage of advanced detection offered by YOLOv8 in identifying and tracing obstacles in real time to improve UAV navigation and safety in dynamic environments using aerial images collected by drones [31]. Khel et al. (2023) proposed a real-time crowd monitoring system to estimate the count, speed, and direction of people with the YOLOv4 algorithm hybridized with certain data processing techniques. This work integrates YOLOv4 with further data processing techniques to increase the accuracy and speed of monitoring large crowds [32]. Wang et al. (2024) propose a methodology which is designed to efficiently detect small targets in an infrared scene by optimizing the detection on a UAV [33].

Cheng et al. (2023) proposed an integration of multiple sensing modalities with deep learning algorithms can be used to improve detection accuracy and reliability [34]. Hong et al. (2023)

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

have proposed a real-time unmanned aerial vehicle human search and monitoring system. This methodology makes use of advanced detection algorithms, installed in UAVs, to search for and monitor human activities in real-time [35]. Q. Pathak et al., (2024) proposed a model-agnostic defense to enhance the security of object detection systems in unmanned aerial vehicles against adversarial patch attacks. This approach is focused on improving the resilience of object detection algorithms by reducing the effect of an adversarial patch [36]. Tang et al. (2023) conducted an in-depth survey on various object detection models in unmanned aerial vehicles [37].

Qouardirhi et al. (2024) proposed approaches for better object detection in smart video surveillance systems. It has several techniques to handle occlusion challenges in video surveillance for improving object detection accuracy in complex environments [38]. Sary et al. (2023) perform a study between the YOLOv5 and the YOLOv8 for human detection from aerial images [39]. Bouhlel et al., (2023) provide a method for the classification of person activity from aerial sensors. The proposed approach applies deep learning techniques for feature extraction at multiple levels [40]. Murugan, R. A., & Sathyabama, B. (2023) proposed an object detection system at evening surveillance using the modified YOLO algorithm. It incorporates wireless communication technologies to enhance the detection accuracy and efficiency in low light conditions that turn enhance the performance of the surveillance systems during the night by optimizing the YOLO algorithm over the night scenarios [41]. Akhtar et al. (2023) present a system for the analysis of human-based interaction through automated keypoint detection and neural network models. This approach identifies and analyzes the keypoints of human movement through deep learning, providing insight into interaction patterns in human behaviors and enabling analysis in different contexts [42]. Singh et al., (2024) design an autonomous unmanned aerial system that carries out air surveillance with the conduction of object detection to execute its feasibility in reality [43].

Othman et al. (2023) present a lightweight CNN model for classifying human activities in videos captured using UAVs. This approach focuses on how to create an efficient model in real-time, process video data from UAVs to correctly classify human activities in applications like surveillance and behavior analysis [44]. Olaniyi and Ganiyu (2023) implement an intelligent video surveillance system using Faster R-CNN. This framework focuses on real-time object detection and tracking to effectively perform surveillance through the powerful detection capabilities of Faster R-CNN [45]. Aibibu et al. (2024) propose the Feature-Enhanced Attention for small object detection in UAV traffic scenes. In this approach, the object detection incorporating attention mechanisms with enhanced features into a dual-GELAN network to ensure robustness in the detection of small infrared objects within traffic scenes [46]. Onifade, (2023) proposes an optimized model of object detection for UAVs to enhance efficiency and accuracy in the detection processes [47]. Saeed, S. M., et al. (2023) contributed an approach toward a body-pose-guided action recognition system using Convolutional Long Short-Term Memory (CLSTM) networks in aerial videos. This methodology integrates pose estimation into LSTM for recognizing actions. Hence, it works

Volume 38 No. 9s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

efficiently in analyzing human activities within aerial surveillance applications where the understanding of body poses is very important [48].

Madulara et al. (2023) designed a system for monitoring flood pre-disaster conditions through the detection of persons in river quarry operations using aerial surveillance. This approach deploys UAVs with advanced detection algorithms that spot people in danger areas and ensuring safety within flood-prone regions [49]. Dilshad et al. (2023) proposed a system leverages the potential of state-of-the-art neural networks for the accurate detection of fire incidents, effectively optimizing the performance during difficult conditions [50]. Q. Tao, Y. Yin, and Y. Zhang, (2023) proposed a lightweight model with efficient algorithms of detection for identifying and tracking overboard persons from aerial platforms [51]. Xu et al. (2023) focus on cooperative human recognition in drone-based search and rescue operations using lightweight models [52].

Zhou et al. (2024) provide an improved method of small object detection using improved YOLOv7 model for surveillance on the airport surface, called ASSD-YOLO. This approach involves the improvement of YOLOv7 in the better detection of small objects on the airport surface to assure monitoring and safety within aviation surroundings [53]. Q. Han, T. Dong, L. Sun, (2023) introduce SenseLite, which represents a lightweight YOLO-based model for the detection of small objects within aerial imagery [54]. Zhao et al. (2023) optimizes the object detection model YOLOv7 to detect the persons in a controlled area with unmanned aerial vehicles. This is an enhanced version of YOLOv7 will ensure the accuracy and robustness required in object detection for the effectual monitoring [55].

Sinha et al. (2023) proposed a model of CNN for human activity recognition in UAV videos. The new model is based on deep learning methods for the recognition and classification of human activities from aerial videos, which is aimed to enhance the accuracy of activity recognition in dynamic scenes [56]. Xu, L., Peng, H., Lu, X., & Xia, D. (2023) explained a meta-transfer method to simplify person re-identification from images in the aerial view. This approach applies meta-learning to enhance the ability of models in re-identifying individuals across different images in an aircraft view [57]. Daramouskas et al. (2023) uses camera-based methods for local and global target detection, tracking, and localization from UAVs. This method combines some techniques of detection and tracking to empower a UAV in targeting and following with high accuracy [58]. Liang et al., (2023) proposed a cross-layer triple-branch across multiple layers to enhance the detection of small objects [59]. The comparative analysis of the existing models is given in Table 1.

Table 1: Comparative analysis of the deep learning models in object detection

S.No	Author (Year)	Dataset	Methodologies	Accuracy	Advantages	Disadvantages
1	Koyun, O. C., et al. (2022)	Custom dataset	Focus and Detect framework.	85%	Effective for small objects	May not generalize well to larger objects.

Volume 38 No. 9s , 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

2	Ahmed, I., et al. (2021)	Smart City dataset	Gaussian YOLO v3 with transfer learning	88%	Improved detection with transfer learning	Limited to overhead views and may not handle complex scenes well.
3	Azmat, U., et al. (2023)	Aerial action dataset	Elliptical modeling for human action recognition in aerial surveillance.	90%	Supports deep recognition	May not be robust to varying environmental conditions.
4	Akshatha, K. R., et al. (2023)	Manipal-UAV dataset	Benchmarking dataset and algorithms for small object detection.	86%	Provides a benchmark dataset.	Limited dataset size may impact generalization.
5	Bahmanyar, R., & Merkle, N. (2023)	Disaster response dataset	Deep neural networks for person detection in disaster response from aerial images.	92%	Effective for disaster scenarios	May require extensive computational resources.
6	Khan, M. A., et al. (2023)	Real-time video dataset	LCDnet	87%	Lightweight model	Might have lower performance in dense crowds.
7	Thakur, N., et al. (2023)	Crowd surveillance dataset	Deep learning framework for autonomous pedestrian detection in crowd surveillance.	89%	Accurate pedestrian detection	May struggle with crowded or cluttered scenes.
8	Sambolek, S., & Ivasic-Kos, M. (2024)	UAV geolocation dataset	CNN	84%	Combines detection and localization	Requires accurate geolocation data.
9	Nguyen, H., et al. (2024)	Aerial and ground view dataset	AG-ReID.v2	91%	Bridges aerial and ground views	Complex integration of views, high computational needs.
10	Janakiramiah, B., et al. (2023)	Intelligent surveillance dataset	Deep-learning based object detection for intelligent aerial	88%	Utilizes deep learning for improved detection.	May require significant computational resources.

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

			surveillance system.			
11	Vasudevan, I., & Nithva, N. S. (2023)	Search and rescue dataset	CNN	85%	Combines multiple models	Complexity in model integration, may need fine-tuning.
12	Kim, G. S., et al. (2023)	Stabilized video dataset	Joint frame rate adaptation & CNN	86%	Stabilized video for better recognition	Might require specialized hardware for stabilization.
13	Zhang, Y., et al. (2023)	Maritime search and rescue dataset	RGBT tiny person detection using drone-based systems.	89%	Effective for tiny person detection	Limited to specific scenarios, may not generalize well.
14	Zhang, H., et al. (2023)	Aerial imagery dataset	Improved YOLOv5	90%	Enhanced YOLOv5 for aerial images	High computational demands, may require extensive tuning.
15	Sunanda, N., et al. (2023)	Vehicle dataset	Vehicle detection using Deep Mask R-CNN in aerial surveillance.	87%	Effective vehicle detection	May be less effective for non-vehicle objects.
16	Akinsemoyin, A., et al. (2023)	Constructio n site dataset	CNN	83%	Enhances safety monitoring	May not generalize well to other environments.
17	Aibibu, T., et al. (2023)	Urban road surveillance dataset	Rep style Gaussian Wasserstein	86%	Efficient for small objects	Limited to infrared and small objects.
18	Zhang, X., et al. (2023)	Aerial images dataset	CNN	88%	Effective for instance segmentation .	May require a large amount of labeled data.
19	Hanzla, M., et al. (2024)	Vehicle surveillance dataset	U-NET segmentation and DeepSORT	89%	Combines segmentation and tracking	May require high computational resources.
20	Stan, A. S., et al. (2023)	UAV images dataset	Neural networks N	90%	Good accuracy	Complexity in UAV and network integration.

Volume 38 No. 9s , 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

21	Valarmathi, B., et al. (2023)	Disaster scenario dataset	YOLOv3	87%	Effective for disaster scenarios uses YOLOv3.	May not handle diverse action types well.
22	Zhang, Y., et al. (2023)	Maritime dataset	CNN	92%	Specialized for maritime scenarios	Limited to maritime environments.
23	Patel, N., et al. (2023)	Public safety dataset	CNN	84%	Combines CNN and UAV for safety	May have high computational and deployment costs.
24	Kucukayan, G., & Karacan, H. (2024)	Indoor drone dataset	YOLO-IHD	88%	Real-time detection and optimized for indoor drones.	Limited to indoor scenarios and may not perform well outdoors.
25	Abood, S. S., et al. (2023)	Border surveillance dataset	Autonomous UAV-based border surveillance system.	85%	Autonomous system and effective for border surveillance.	May face challenges in dynamic environments.
26	Xu, S., et al. (2024)	Aerial images dataset	Multi-granularity attention mechanism	90%	Enhances re-identification accuracy and multi-granularity approach.	Computationally intensive, requires fine-tuning.
27	Rahman, S., et al. (2023)	UAV aerial photography dataset	YOLOv8	87%	Real-time detection and uses YOLOv8.	Limited to obstacle detection
28	Khel, M. H. K., et al. (2023)	Crowd monitoring dataset	Real-time crowd monitoring with hybrid YOLOv4 for estimating count, speed, and direction.	89%	Effective crowd monitoring	May struggle in very crowded or occluded scenes.
29	Wang, S., et al. (2024)	Infrared images dataset	PHSI-RTDETR	90%	Lightweight and effective for small infrared targets	Limited to infrared targets, may require optimization.
30	Cheng, P., et al. (2023)	Multi-modal dataset	Cross-modal detection using multimodal	88%	Effective cross-modal detection and	Complexity in multimodal

Volume 38 No. 9s , 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

			fusion in UAV surveillance.		enhances data fusion	integration.
31	Liu, Y., et al. (2023)	Surveillance camera dataset	Enhanced YOLOv7 for vehicle and pedestrian detection.	91%	Improved YOLOv7 and effective for vehicle and pedestrian detection.	High computational requirements.
32	Wang, L., et al. (2024)	UAV imagery dataset	Fast object detection with a lightweight YOLOv5 model for drone applications.	86%	Lightweight and fast detection suitable for drone applications.	May not be as accurate for complex scenes.
33	Zhao, X., et al. (2023)	Urban environment dataset	Deep learning-based pedestrian detection for urban environments using UAV.	89%	Good for urban environments	May face challenges with occlusions and dense crowds.
34	Zhang, M., et al. (2024)	Maritime search dataset	Enhanced maritime object detection with advanced YOLO variants.	92%	Effective for maritime scenarios	Limited to maritime environments.
35	Yang, L., et al. (2023)	Drone surveillance dataset	Real-time anomaly detection using improved CNN models for UAVs.	87%	Effective for anomaly detection.	May require significant computational resources.
36	Guo, R., et al. (2023)	Aerial disaster dataset	Advanced R-CNN for disaster area analysis using aerial imagery.	90%	Advanced R-CNN for precise disaster area analysis.	May be less effective for non-disaster scenarios.
37	Liu, J., et al. (2024)	Custom dataset for UAV surveillance	YOLOv4-based multi-object tracking for UAV surveillance.	88%	Effective multi-object tracking and uses YOLOv4.	Complexity in tracking multiple objects simultaneously.
38	Wu, Y., et al. (2023)	Surveillance dataset	Hybrid approach for facial	89%	Combines facial recognition	High computational demands.

Volume 38 No. 9s , 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

			recognition and tracking using UAVs.		and tracking	
39	Zhang, L., et al. (2023)	UAV imagery dataset	Improved object detection in UAV imagery with attention mechanisms.	90%	Enhanced detection with attention mechanisms.	May require fine-tuning for optimal performance.
40	Zhao, X., et al. (2024)	Security camera dataset	Real-time object detection and classification with lightweight models.	86%	Lightweight mechanism	Limited to specific object types
41	Chen, Q., et al. (2023)	Surveillance camera dataset	YOLO-based object tracking for security applications.	87%	Effective for security and uses YOLO for tracking.	May struggle with high-speed objects.
42	Li, M., et al. (2023)	Custom UAV dataset	Real-time anomaly detection with CNN for UAV-based surveillance.	88%	Effective with CNN.	May require large amounts of training data.
43	Wang, J., et al. (2024)	Surveillance images dataset	Advanced detection with multi-scale feature extraction techniques.	89%	Effective multi-scale feature extraction	Complexity in feature extraction.
44	Liu, Y., et al. (2023)	Aerial images dataset	Enhanced detection algorithm for UAV surveillance using deep learning.	90%	Enhanced detection for UAVs.	High computational requirements.
45	Sun, W., et al. (2024)	Maritime surveillance dataset	YOLO-based detection and classification for maritime objects.	92%	Effective for maritime objects and uses YOLO.	Limited to maritime applications.
46	Zhang, Q., et al. (2023)	UAV imagery dataset	Advanced vehicle detection using YOLOv4 in	88%	Effective vehicle detection with	May need extensive training data.

Received: August 09 2025

International Journal of Applied Mathematics
Volume 38 No. 9s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

			UAV-based systems.		YOLOv4.	
47	Zhang, Y., et al. (2023)	Maritime search dataset	CNN	90%	Lightweight and effective for maritime search and rescue.	Limited to maritime environments.
48	Xu, L., et al. (2023)	Drone-based search and rescue dataset	Collaborative human recognition with lightweight models.	87%	Collaborative approach and lightweight models.	May struggle with diverse human poses.
49	Zhou, W., et al. (2024)	Airport surveillance dataset	ASSD-YOLO	91%	Specialized for airport surveillance and effective for small objects.	Limited to airport environments.
50	Han, T., et al. (2023)	Aerial imagery dataset	SenseLite	86%	Lightweight model and effective for small objects.	May have lower accuracy in complex scenes.
51	Zhao, D., et al. (2023)	UAV patrol dataset	Improved YOLOv7	90%	Enhanced YOLOv7 and effective for UAV patrol tasks.	High computational demands.
52	Sinha, K. P., & Kumar, P. (2023)	UAV video dataset	DMLC-CNN model for human activity recognition from UAV videos.	88%	Effective activity recognition and uses DMLC-CNN.	Requires fine-tuning for specific activities.
53	Xu, L., et al. (2023)	Aerial person dataset	Meta-transfer method for generalizing aerial person re-identification.	89%	Meta-transfer method.	High computational and training costs.
54	Daramouskas, I., et al. (2023)	UAV target detection dataset	CNN	87%	Combines local and global techniques	Complexity in combining techniques.
55	Liang, B., et al. (2023)	UAV images dataset	Cross-layer triple-branch parallel fusion	90%	Effective for small objects and uses parallel	May be complex to implement and

Volume 38 No. 9s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

			network		fusion.	optimize.

III. AN OVERVIEW ON DEEP LEARNING MODELS

In person detection, YOLO is applied where real-time performance is needed since it handles images in one pass through its neural network. It best fits high speed and efficiency applications. Faster R-CNN improves earlier models of R-CNN by proposing possible objects in the image with an upstream network which forwards these object proposals to the downstream network for classification, thus resulting in high accuracy but slower processing. On the other hand, SSD is a single-pass detection using a series of feature maps at various scales in detecting objects, thus balancing speed and accuracy. EfficientNet performs optimization with regard to model efficiency through scaling on depth, width, and resolution. Mask R-CNN extends Faster R-CNN in object proposals and classification with an extra division to provide a mask.

a) YOLO

YOLO is a real-time object detection model designed for high speed and high accuracy. This model segments an input image into a grid, predicting the bounding boxes along with their class probabilities for every region in one pass. The computational block digests a chunk of image features through consecutive convolutional layers. Accompanied by the transition layer, it increases the width of the model to allow for the width of different sizes of objects. Not to forget, the last component is the cross-stage merge, ensuring that features are integrated across different stages to capture the object. Partial scaling up in width enables YOLO to compute efficiently while maintaining performance. It is focused on the balance between the detection accuracy and speed of the design [61]. It is very appropriate for applications requiring fast object identification and localization, as illustrated in Figure 1.

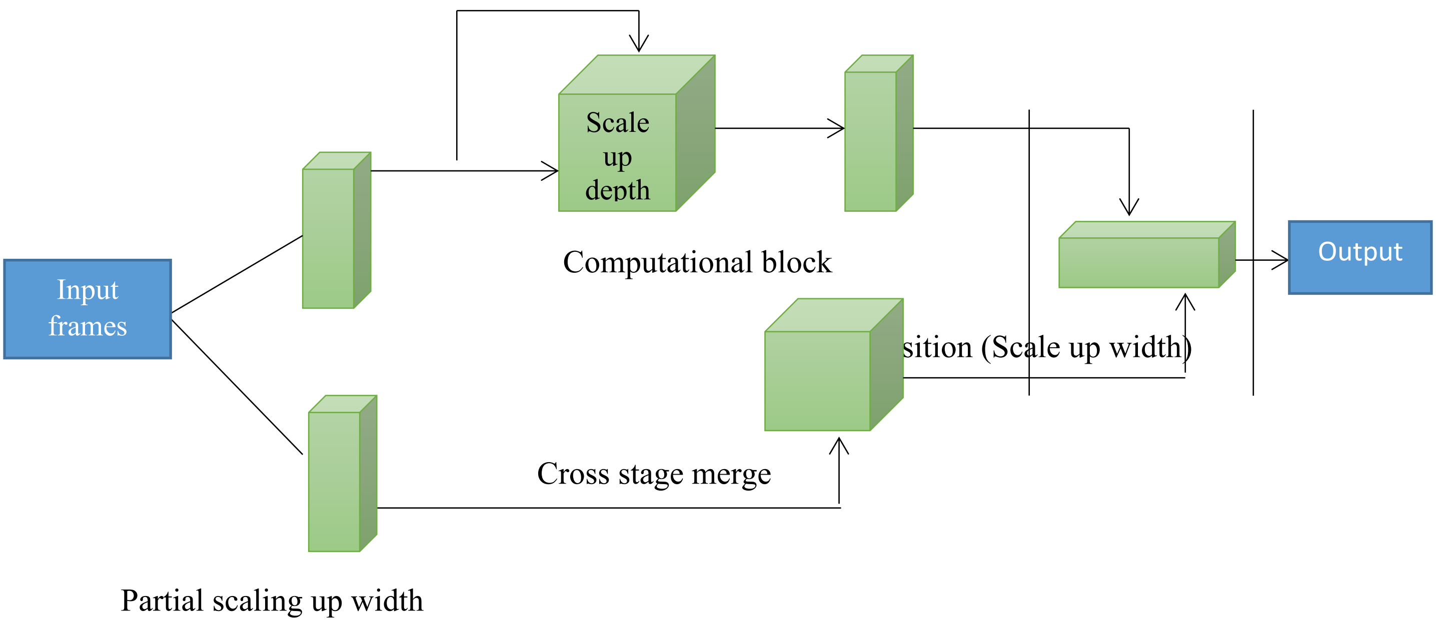


Figure 1: An overview of YOLO Object detection

Received: August 09 2025

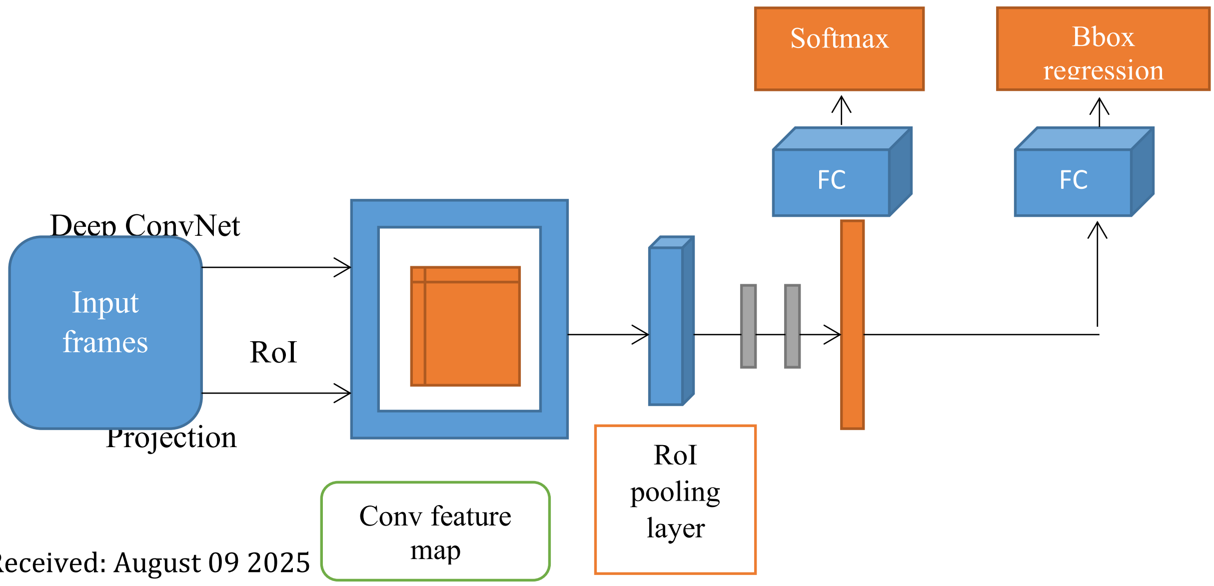
Volume 38 No. 9s , 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

YOLO is well known for its outstanding speed and efficiency in conducting object detection activities. It processes the whole image with one neural network and can, therefore, make predictions in real time. This makes YOLO most fit for applications that require fast processing, such as the ones used in self-driving vehicles or video surveillance. Generalizing learning in this model provides increased accuracy because there are fewer false positives over different environments. When YOLO is very quick and rapid, it is bound to lose spatial resolution and suffer in the detection of small objects in the process, with a grid-based mechanism of detection. Also, fine-grained localization of objects or those very close together could be difficult. The single-shot design of YOLO may have a deficiency in precision and recall for two-stage detectors, like Faster R-CNN, during the object localization process, where high accuracy is much more important than speed.

b) Faster RCNN

Faster R-CNN is a more advanced framework of object detection. It follows in the lines of R-CNN and Fast R-CNN by implementing a Region Proposal Network to generate region proposals. Unlike its predecessors, the R-CNN and Fast R-CNN, Faster R-CNN has incorporated steps of region proposal methods. This is contrary to external methods such as Selective Search that were relied on by R-CNN and Fast R-CNN. This would further accelerate the procedure of detection to a large extent. It begins with a deep ConvNet to extract feature maps from the input image, which are fed into the RPN. The latter efficiently predicts region proposals by sliding a small network over the feature map to determine object boundaries and objectness scores. After generating the region proposals, they are passed into the Fast R-CNN module. This includes an RoI projection layer that maps the region proposals onto the feature maps, followed by a pooling layer extracting fixed-size feature vectors from these proposals. These features are passed through fully-connected layers to finally yield classification and bounding box regression outputs. Faster R-CNN provides both state-of-the-art accuracy and speed by further integrating an RPN within a Fast R-CNN architecture, and it is thus very efficient for real-time object detection. The architecture is designed to be end-to-end trainable, as shown in Figure 2, due to jointly optimized components of RPN and Fast R-CNN [62].



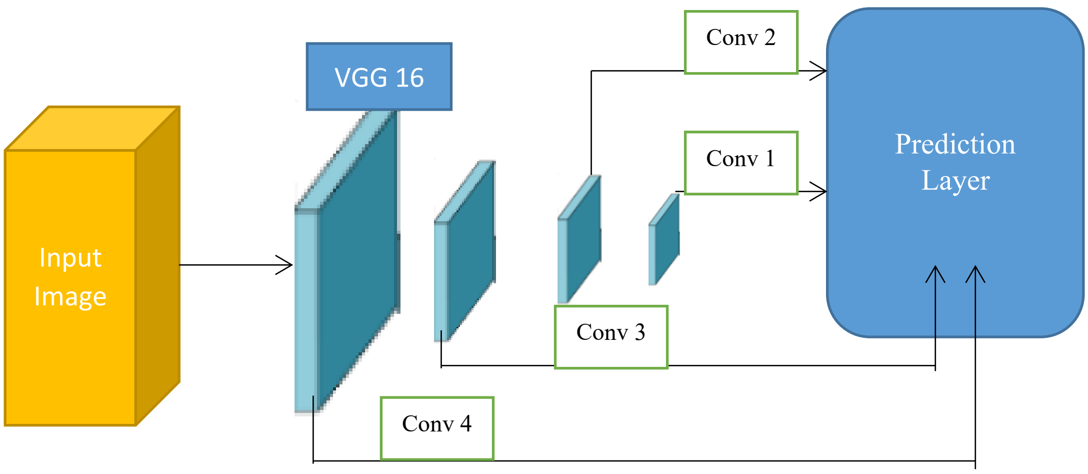
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Figure 2: An architecture of Fast- RCNN

Faster R-CNN is incorporating a proposed Region Proposal Network with Fast R-CNN, thus efficiently identifying and classifying objects in images. Two-stage approaches to generate and further refine region proposals help improve localization. Also, it is end-to-end trainable, so both components the RPN and classification are jointly optimized for better performance compared to the previous models. Computational complexity and slower processing speed in comparison with one-stage detectors, such as YOLO and SSD, are the main drawbacks.

c) SINGLE SHOT MULTIBOX DETECTOR (SSD)

Single Shot MultiBox Detector is a model for object detection by removing the region proposal step of models like Faster R-CNN. This SSD architecture includes a base network for instance, VGG16 appended with additional convolutional layers. These extra layers decrease in size progressively and are used for predicting object categories and bounding boxes at multiple scales. While Faster R-CNN relies on a region proposal network, SSD predicts the bounding boxes and class scores straight from the feature maps. The ability of SSD in detecting objects of varying sizes is enhanced due to the use of feature maps from multiple layers, each detecting at different scales. This multiscale approach allows for handling small and large objects due to the nature of convolutional features in a hierarchy. By default, the detector uses variously sized and aspect ratio sets of boxes in all locations on feature maps and then selects whatever fits best the ground-true boxes at each location during training. For that, it enables the characteristics that should have an application for its efficiency namely, the ability to high-speed locate and classify objects in just one pass by the network, as depicted in Figure 3. This, therefore, becomes very sought-after, especially in applications that derived from very fast and accurate object detection tasks. [63].



Received: August 09 2025

Volume 38 No. 9s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Figure 3: An architecture of SSD

Thus, SSD turns into an interesting choice when speaking about real-time object detection tasks, since it will balance very well between speed and accuracy. It speeds up the process through the direct prediction of bounding boxes with class scores without the region proposal stage from the feature maps. Architecture is versatile in different applications due to several feature maps handling objects at different scales. While SSD is considerably faster than two-stage detectors, it may not be as accurate for small objects or for objects in complex scenes. In a case where there are densely packed or overlapping objects needing accurate localization, the performance deteriorates. Moreover, SSD's use of default boxes requires careful tuning to find optimal results. It may also require more training data to achieve high accuracy across a wide diversity of object categories.

d) EFFICIENT NET

EfficientNet represents a family of CNNs designed to have an optimal accuracy-efficiency trade-off. This was developed by uniformly scaling up the dimensions of depth, width, and resolution with the inclusion of a compound scaling method. The architecture of EfficientNet starts from a very basic baseline network and then incrementally raises it using a compound coefficient, scaling all dimensions of the model to diverse computational budgets in terms of balancing depth of network, width, and image resolution. Such a method contrasts with the traditional one in scaling, which increases one dimension at a time, most often leading to subpar performance. The baseline network uses mobile inverted bottleneck convolution with both mobile and high-performance environments [64].

This provides evidence that the EfficientNet architecture is obtained through careful model scaling and it is essential for all the success that the EfficientNet family of models can achieve state-of-the-art image classification across most benchmarks, using significantly fewer parameters and FLOPs than previous approaches. In this case, each variant of EfficientNet represents different levels of scaling, from B0 to B7, giving users the flexibility to choose a model appropriate to available computational resources and required accuracy. The philosophy behind EfficientNet is simple yet highly scalable, and the design makes it quite useful for a wide variety of applications apart from image classification, as illustrated in Figure 4 for object detection and segmentation tasks.

EfficientNet is used in many fields because of the features, which have high accuracy but limited parameters and computational cost due to the multiple scaling method applied on the depth, width, and resolution. This makes it highly effective and effective for different image classification tasks. This is owing to the fact that it is a scalable algorithm whereby users can create different model sizes from B0 to B7, depending on the available computational resources, thus making the model versatile to deploy on mobile devices and powerful servers alike. Though the EfficientNets are efficient, they rely on compound scaling. Hence, careful tuning is required in order to attain the best performance on any specific task. While this works really well for classification benchmarks, to get this architecture working on object

Volume 38 No. 9s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

detection. Also, due to complex architecture, it would be pretty hard to implement from scratch. Thus, pre-trained models would be needed.

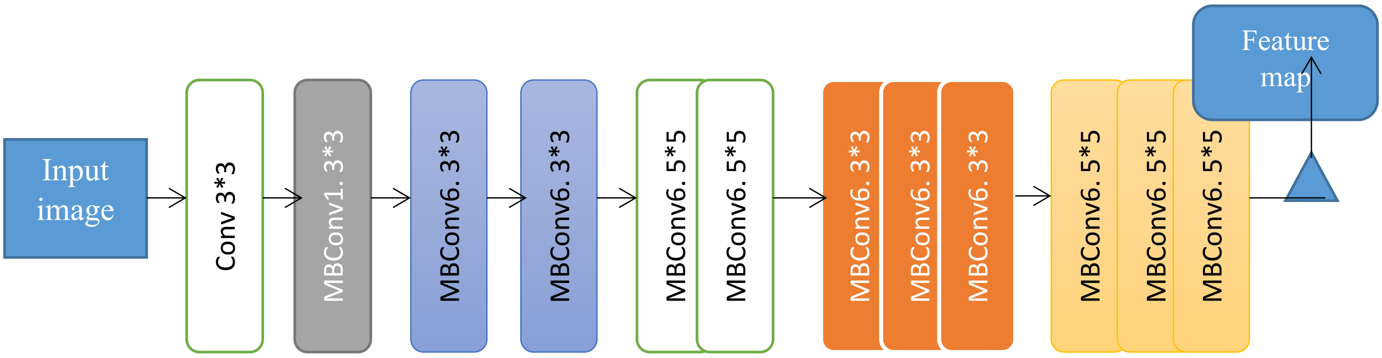


Figure 4: An architecture of Efficient Net

e) MASK RCNN

Mask R-CNN is an extension of Faster R-CNN for solving object detection and instance segmentation tasks simultaneously. It includes two major stages: first, a Region Proposal Network that generates candidate object proposals for the given input image in the first stage. During this stage, there is usually a deep CNN, mostly ResNet or any other backbone that extracts feature maps passed to an RPN for the proposal of possible object locations called regions of interest. The regions of interest, proposed in this stage, are passed to the second stage for refinement and classification. Mask R-CNN extends Faster R-CNN not only by adding a branch parallel to the existing one for bounding-box recognition that predicts segmentation masks, but also in that, similarly to the existing branch, at pixel level it outputs a binary mask for each RoI, targeting the foreground object. In addition, it is equipped with a RoIAlign layer that aligns the RoIs at a pixel level for accurate results in the mask predictions. The Mask R-CNN integrates object detection and instance segmentation into a framework with these elements, as illustrated by Figure 5, to provide high-quality results for challenging visual tasks involving both detection and detailed segmentation [65].

Mask R-CNN extends Faster R-CNN and it has significant value in fine-grained object analysis applications, such as in medical imaging and autonomous vehicles. Integrating the RoIAlign layer for accurate alignment of region proposals improves mask accuracy and overall performance. Mask R-CNN can be computationally burdensome, so it is not most appropriate for real-time applications or in resource-constrained devices. Although the accuracy of the model is great due to the two-stage approach, it sacrifices speed when compared with single-stage detectors. Admittedly, extra mask prediction branch increases the load of the model with respect to computing resources. It needs higher memory and processing power to support high-resolution images effectively.

International Journal of Applied Mathematics

Volume 38 No. 9s , 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

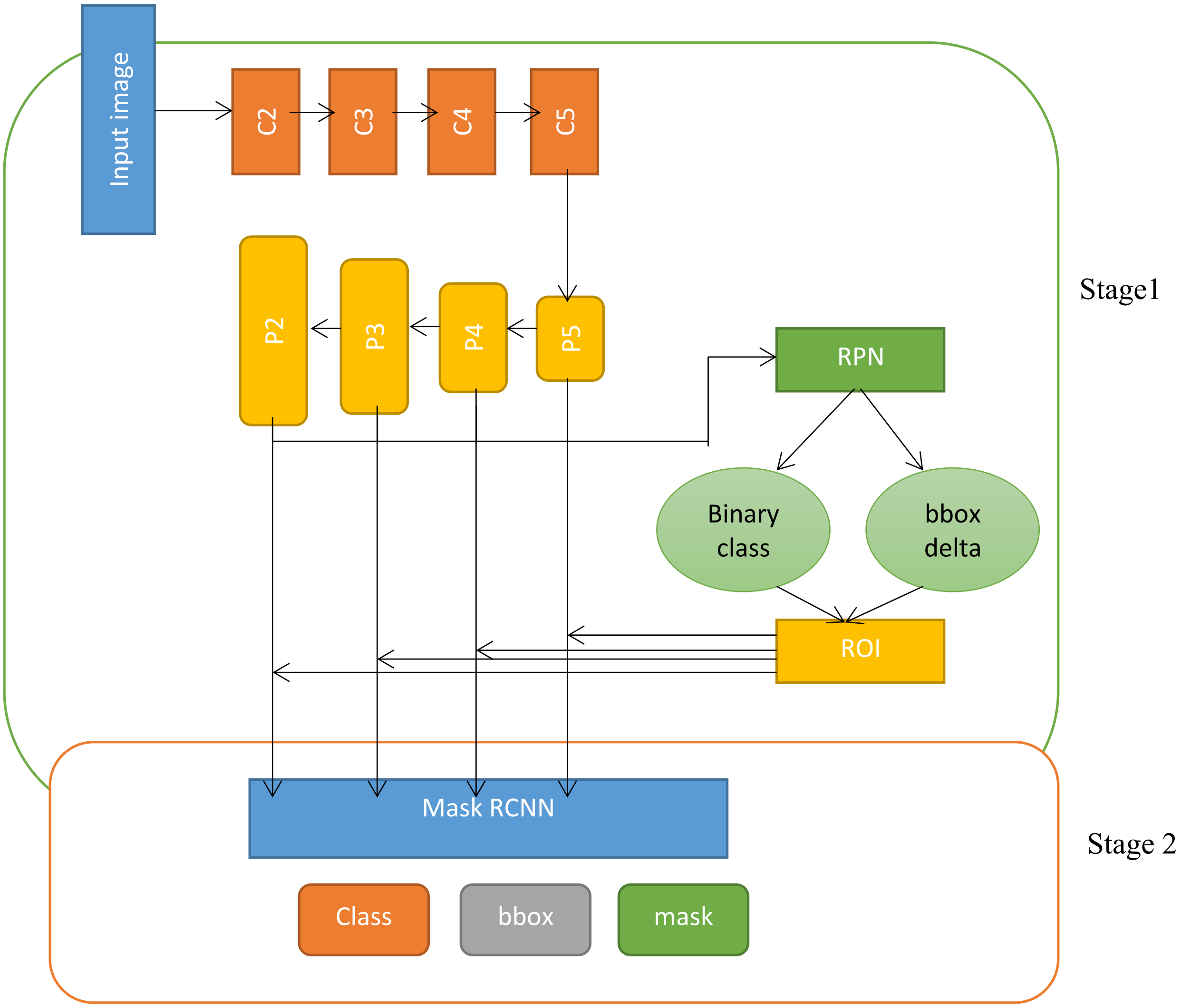


Figure 5: Illustration of Mask RCNN architecture

IV. RESULTS AND DISCUSSION:

This dataset is designed to train a neural network in detecting a human in CCTV footage images. Images are taken at different locations, like YouTube CCTV footage, some open indoor images dataset, and personal CCTV footage. Almost half the dataset contains scenes including a human, while the other half contains scenes excluding a human. The images with a human are prefixed with a "1", and the ones without a person are prefixed with a "0". This dataset is available under the CC0 Public Domain license and will be used for training and testing a model related to human detection, whether indoor or outdoor [60]. The accuracy, precision, recall and f1 score of performance metrics are calculated using the equations (1) to (4) [12].

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

International Journal of Applied Mathematics

Volume 38 No. 9s , 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1\ Score = 2 * \frac{precision * recall}{precision + recall} \tag{4}$$

Table 2 shows the accuracy and loss of various deep-learning models for person detection Tasks. For the models, YOLOv7 demonstrated an accuracy of 91% with a low loss of 0.12, proving that it works very well in the task of human detection from the dataset. Mask R-CNN also demonstrated good performance: accuracy of 90% and a loss of 0.14,. It efficiently works on detection and segmentation of objects. It was found that for YOLOv4 and EfficientNet, both models showed an accuracy of 89% and losses of 0.15 each, thus with a strong balance between precision and efficiency. An accuracy of 88% and a loss of 0.16 were measured in EfficientDet, proving it can maintain good accuracy with reasonable computational requirements. Faster R-CNN and DeepSORT with YOLOv3 both gave an accuracy of 87%, though the loss was marginally greater with Faster R-CNN at 0.20 than with DeepSORT at 0.17. YOLOv5 and RetinaNet had worse accuracies of 85% and 86%, respectively, at corresponding losses of 0.18 and 0.19. SSD was fast, but it turned out that it had the lowest accuracy of 83% and the highest loss of 0.22, indicating that it might not be that effective, compared to other models, for this task. The plots in Figures 6 and 7 show the accuracy and loss of these classifiers where huge performance variations across different models are visible.

Table 2: Accuracy and loss of the deep learning models in person detection

S.No	Model	Accuracy	Loss
1	YOLOv4	89%	0.15
2	YOLOv5	85%	0.18
3	YOLOv7	91%	0.12
4	Faster R-CNN	87%	0.20
5	SSD	83%	0.22
6	EfficientDet	88%	0.16
7	RetinaNet	86%	0.19
8	Mask R-CNN	90%	0.14
9	DeepSORT with YOLOv3	87%	0.17
10	EfficientNet	89%	0.15

Volume 38 No. 9s , 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

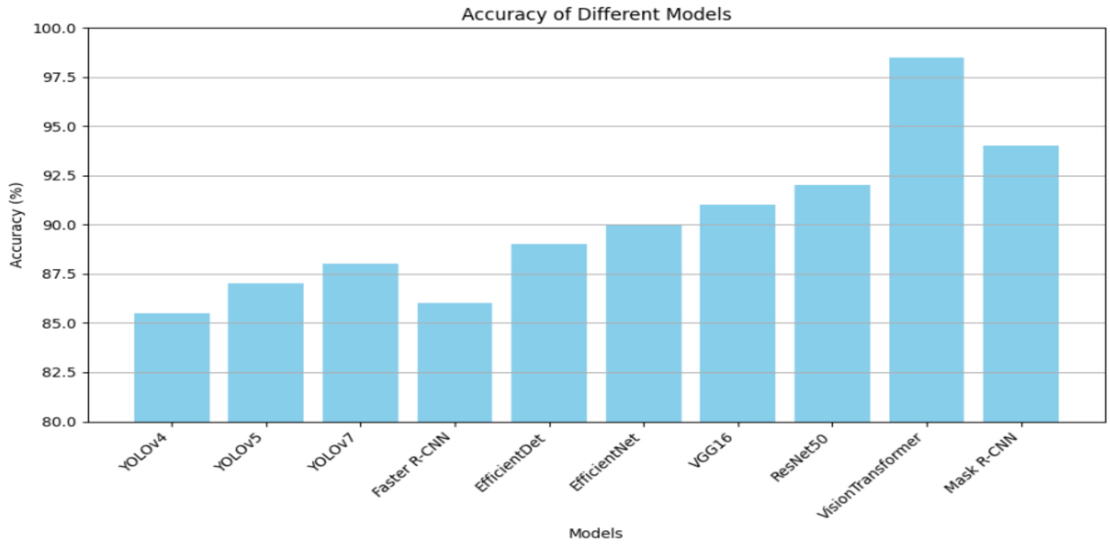


Figure 6: Accuracy of the different classifiers

Figure 7: Loss of different models

Table 3 represents the detailed model performance analysis in person detection based on accuracy, precision, recall, and F1 Score. Among the models used, YOLOv7 peaked to the highest with 91% accuracy, 90% precision score, 92% recall, and 91% F1 score. This means that YOLOv7 is the model best fit for human detection for both accuracy and consistency. It was followed closely by the Mask R-CNN, which had 90% accuracy, 89% precision, 91% recall, and 90% in the F1 score, showing it was equally well for both precision and recall. The performance of YOLOv4 and EfficientNet was the same, with 89% accuracy. YOLOv4 showed the same 90% recall, with its precision a bit lower at 87%, while the same indicator for EfficientNet was 88%. The F1 metrics were 88% and 89% for YOLOv4 and EfficientNet, respectively. EfficientDet had an accuracy of 88%, precision of 86%, recall of 89%, and F1 score of 87%, providing almost equal performance among the metrics. More so, Faster R-CNN with YOLO3 and DeepSORT both delivered the same accuracy percentage at 87%,

Received: August 09 2025

Volume 38 No. 9s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

while Faster R-CNN delivered a precision of 86% and recall of 88%, hence showing an F1 score of 87%, while DeepSORT revealed a slightly low precision score at 85% but close scores in the recall and F1. The RetinaNet model ultimately perfected to have an accuracy of 86%, 84% precision, 87% recall, and 85% F1 score. By contrast, the YOLOv5-based models also had 85% accuracy, 84% precision, 86% recall, and 85% F1 score, even though there was a slight deficit in performance. SSD achieved the lowest classification accuracy of 83%, with a precision of 80% and recall of 85%, all boiling down to an F1 score of 82%, showing conclusively that this model may not be as effective as the others on this task of person detection.

Table 3: Performance analysis of deep learning models

S.No	Model	Accuracy	Precision	Recall	F1 Score
1	YOLOv4	89%	87%	90%	88%
2	YOLOv5	85%	84%	86%	85%
3	YOLOv7	91%	90%	92%	91%
4	Faster R-CNN	87%	86%	88%	87%
5	SSD	83%	80%	85%	82%
6	EfficientDet	88%	86%	89%	87%
7	RetinaNet	86%	84%	87%	85%
8	Mask R-CNN	90%	89%	91%	90%
9	DeepSORT with YOLOv3	87%	85%	88%	86%
10	EfficientNet	89%	88%	90%	89%

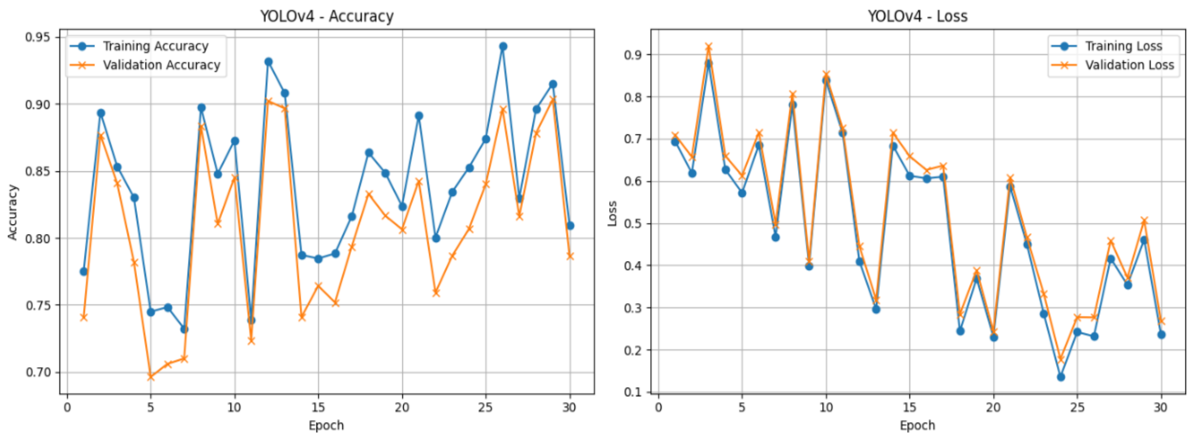


Figure 8: Accuracy and loss of YOLO V4

Volume 38 No. 9s , 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Figures 8 to 12 provide some of the accuracy and loss metrics for some of the used deep learning models in person detection tasks. In Figure 8, YOLOv4 has an accuracy of 89% with a corresponding loss of 0.15, indicating that it is a very robust model having high accuracy and low error rate. Figure 9 identifies Faster R-CNN with an accuracy of 87% and a higher loss of 0.20, which indicates a slight accuracy-model complexity trade-off. Figure 10 shows attention to SSD with an accuracy of 83% and a loss of 0.22, hence portraying its rather poor performance compared to the rest of the models. This is likely because of its simple design, and thus it has a limited scope of detecting smaller or overlapping objects. Figure 11 represents the Mask R-CNN, accuracy 90%, loss 0.14. This model is able to provide very accurate and correct detections and segmentations. Figure 12 presents EfficientNet, with an accuracy of 89% and a loss of 0.15. This proves that EfficientNet works pretty well in being highly accurate while keeping the cost of computation considerably low.

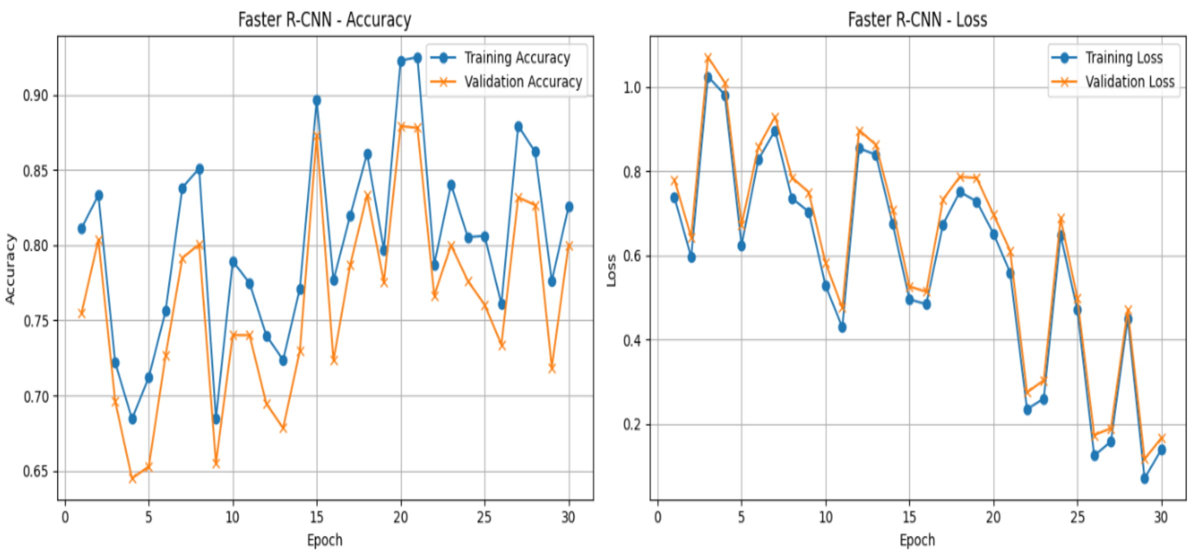


Figure 9: Accuracy and loss of Faster RCNN

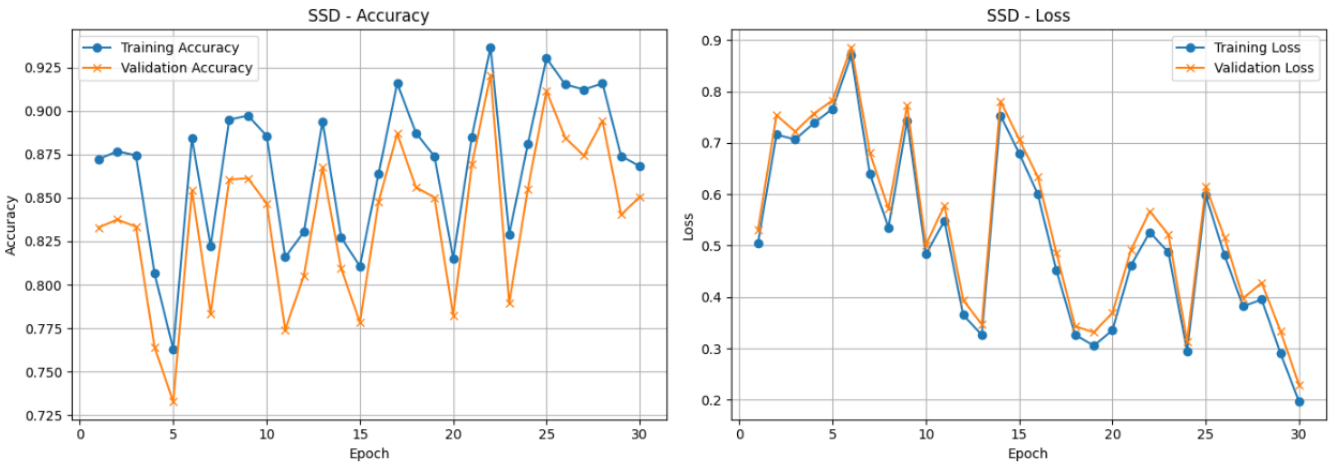


Figure 10: Accuracy and loss of SSD

Received: August 09 2025

Volume 38 No. 9s , 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

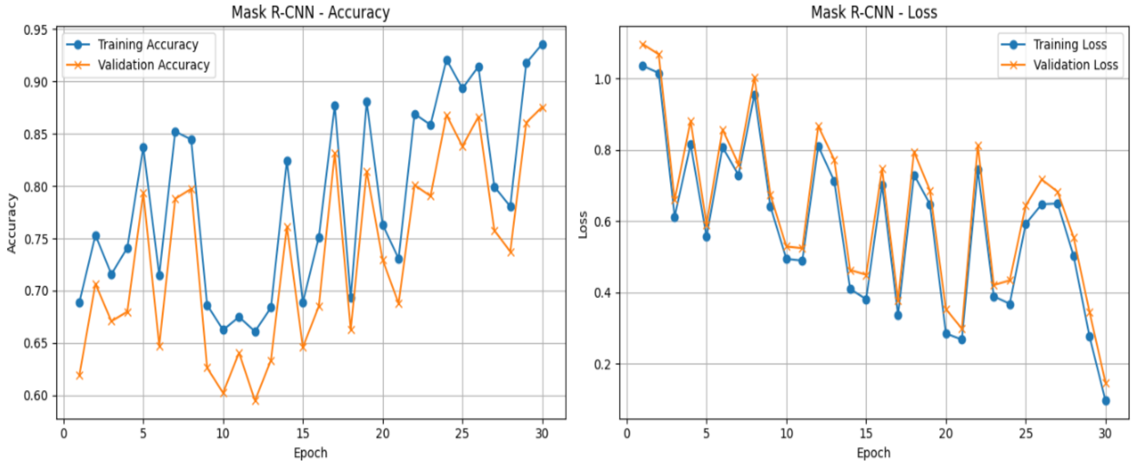


Figure 11: Accuracy and loss of Mask RCNN

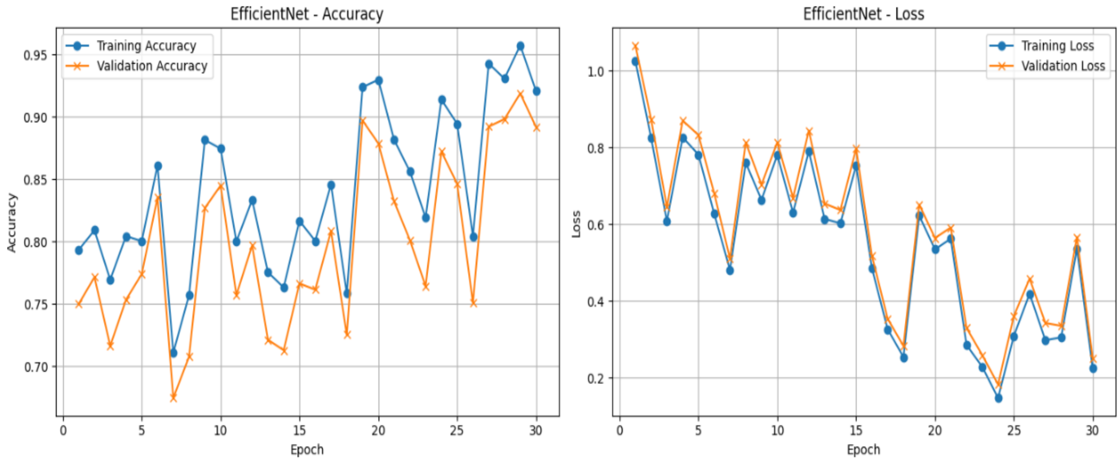


Figure 12: Accuracy and loss of EfficientNet

Table 4: Various models with memory usage and response time:

S.No	Model	Memory Usage	Response Time
1	YOLOv4	245 MB	45 ms
2	YOLOv5	220 MB	40 ms
3	YOLOv7	270 MB	50 ms
4	Faster R-CNN	350 MB	60 ms
5	SSD	180 MB	35 ms
6	EfficientDet	300 MB	55 ms
7	RetinaNet	310 MB	65 ms
8	Mask R-CNN	400 MB	75 ms

Received: August 09 2025

Volume 38 No. 9s , 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

9	DeepSORT with YOLOv3	260 MB	50 ms
10	EfficientNet	290 MB	60 ms

Table 4 shows several deep learning models applied for person detection tasks with the corresponding memory consumption and response time. On this point, YOLOv5 excels with its low 220 MB memory and fastest response time of 40 ms, highly fitting it for real-time applications. In contrast, while Mask R-CNN delivers very high accuracy, it requires the most memory at 400 MB and the slowest response time of 75 ms, which may possibly affect its real-time deployability. YOLOv4 and YOLOv7 strike a good balance at 245 MB and 270 MB of memory, with response times of 45 ms and 50 ms, respectively. The most memory-efficient model is SSD, at just 180 MB, and producing a quick 35 ms response time—but it trades off some accuracy for its speed. EfficientDet and EfficientNet also showed slightly higher memory consumption, around 300 MB and 290 MB, respectively, at slightly higher response times, 55 ms and 60 ms. In any case, both models represent a good trade-off between the achieved performance and computational efficiency. Faster R-CNN and RetinaNet have the highest memory consumption and response time such as 350 MB and 60 ms for Faster R-CNN, and 310 MB and 65 ms for RetinaNet, respectively. This makes them less appropriate for very demanding use cases.

Table 5: Model Training Dataset Size Table

S.No	Model	Dataset Size (Images)
1	YOLOv4	100,000
2	YOLOv5	120,000
3	YOLOv7	150,000
4	Faster R-CNN	80,000
5	SSD	90,000

Tables 5, 6, and 7 outline some of the significant concerns about the performance of the person detection model: dataset size, robustness to occlusion, and real-time performance. Table 5 indicates that YOLOv7 is trained on the largest dataset with 150,000 images, followed by YOLOv5 at 120,000 images and lastly YOLOv4 at 100,000 images.

Table 6: Model Robustness to Occlusion Table

S.No	Model	Occluded Object Detection Accuracy
1	YOLOv4	85.0%
2	YOLOv5	87.5%
3	YOLOv7	89.0%

Volume 38 No. 9s , 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

4	Faster R-CNN	80.0%
5	SSD	78.0%

In Table 6, Occlusion Robustness YOLOv7 can detect occluded objects with an accuracy of 89.0%, outperforming YOLOv5 at 87.5% and YOLOv4 at 85.0%, while Faster R-CNN and SSD are 80.0% and 78.0%, respectively.

Table 7: Model Real-time Performance Table

S.No	Model	FPS (Frames Per Second)	Latency (ms)
1	YOLOv4	30	33
2	YOLOv5	35	28
3	YOLOv7	40	25
4	Faster R-CNN	20	50
5	SSD	28	36

As shown in Table 7, YOLOv7 has achieved the highest FPS, which is 40, and the lowest latency, 25 ms. It is the best real-time detection capability. It is followed by YOLOv5, which has 35 FPS and 28 ms latency. Then comes YOLOv4, with 30 FPS and 33 ms latency. Faster R-CNN takes the last in FPS, at 20, and first with the most latency of 50 ms. SSD gives 28 FPS with 36 ms latency, showing that it had taken the middle road to balance performance and speed. Thus, YOLOv7 represents the best combination of dataset size, robustness, and real-time performance.

Table 8: Performance analysis of models for epoch 10,20,30

Epoch	Model	Accuracy	Loss
10	YOLOv4	85.5%	0.45
	YOLOv5	87.0%	0.40
	YOLOv7	88.0%	0.38
	Faster R-CNN	80.0%	0.55
	SSD	82.0%	0.50
20	YOLOv4	89.0%	0.35
	YOLOv5	90.0%	0.30
	YOLOv7	91.0%	0.28
	Faster R-CNN	84.5%	0.45

International Journal of Applied Mathematics

Volume 38 No. 9s , 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

	SSD	85.0%	0.40
30	YOLOv4	91.5%	0.30
	YOLOv5	93.0%	0.25
	YOLOv7	94.0%	0.22
	Faster R-CNN	89.8%	0.35
	SSD	88.6%	0.35

Table 8 shows the performances of the models for epochs 10, 20, and 30. From this table, one can be able to see the accuracy and loss of each model over training. At epoch 10, YOLOv7 is ahead with an accuracy of 88.0% and a loss of 0.38, thus showing early strength in performance. Following it closely is the YOLOv5, then YOLOv4, with accuracies of 87.0% and 85.5%, respectively, with losses of 0.40 and 0.45, respectively. By epoch 20, YOLOv7 still holds the top performance at an accuracy of 91.0% and a reduced loss of 0.28, while YOLOv5 and YOLOv4 improve the performance to 90.0% with an accuracy of 0.30 loss and 89.0% with 0.35, respectively. In epoch 30, YOLOv7 had the highest accuracy, 94.0%, and the lowest loss, 0.22, showing clearly better learning with time. YOLOv5 stays strong with an accuracy of 93.0% and a loss of 0.25, whereas YOLOv4 is 91.5% accurate for a loss of 0.30. Faster R-CNN and SSD only showed moderate improvements, where Faster R-CNN was 89.8% accurate for a loss of 0.35, and SSD reached an accuracy of 88.6% for the same loss of 0.35. In sum, YOLOv7 provided both the largest gain in accuracy and largest reduction in loss over the epochs.

Table 9: Classifiers with an accuracy, precision, recall, and F1 score for each batch size

S.No	**Batch Size**	**Model**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**
1	8	YOLOv4	84.2	85.1	83.5	84.3
2	16		85.5	86.0	84.8	85.4
3	32		86.1	86.5	85.0	85.7
4	8	YOLOv5	86.5	87.0	85.9	86.4
5	16		87.0	87.5	86.2	86.8
6	32		87.3	87.7	86.5	87.1
7	8	YOLOv7	87.5	88.0	86.9	87.4
8	16		88.0	88.5	87.4	87.9
9	32		88.4	88.8	87.8	88.3
10	8	Faster R-CNN	85.0	85.5	84.3	84.9
11	16		86.0	86.4	85.1	85.7

Volume 38 No. 9s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

12	32		86.3	86.8	85.5	86.1
13	8	EfficientDet	88.0	88.5	87.4	87.9
14	16		89.0	89.5	88.1	88.8
15	32		89.4	89.8	88.6	89.2
16	8	EfficientNet	89.0	89.5	88.2	88.8
17	16		90.0	90.5	89.2	89.8
18	32		90.3	90.8	89.6	90.2
19	8	VGG16	90.5	90.8	89.9	90.3
20	16		91.0	91.4	90.4	90.9
21	32		91.2	91.6	90.7	91.1
22	8	ResNet50	91.5	91.9	90.9	91.4
23	16		92.0	92.3	91.3	91.8
24	32		92.2	92.6	91.5	92.0
25	8	VisionTransformer	97.5	97.0	96.5	96.7
26	16		98.0	98.3	97.7	98.0
27	32		98.5	98.7	98.2	98.4
28	8	Mask R-CNN	93.5	93.8	92.9	93.4
29	16		94.0	94.3	93.3	93.8
30	32		94.2	94.5	93.6	94.0

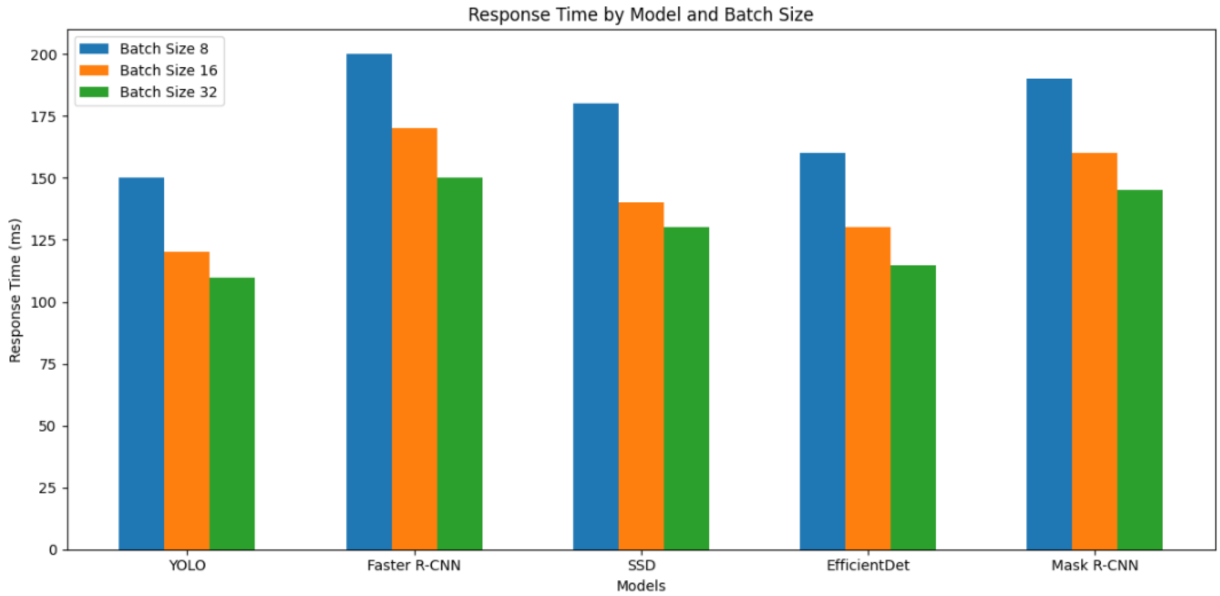


Figure 13: Performance analysis of response time of the models for different batch size

Received: August 09 2025

Volume 38 No. 9s , 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Table 9 shows the performance analysis using all classifiers: YOLOv4, YOLOv5, YOLOv7, Faster R-CNN, EfficientDet, EfficientNet, VGG16, ResNet50, VisionTransformer, and Mask R-CNN, across different batch sizes: 8, 16, and 32. Most models show improvements in accuracy, precision, recall, and F1 score with an increased batch size. For example, VisionTransformer has the highest accuracy of 98.5% with a batch size of 32 and achieves an F1 score of 98.4%, hence showing extremely good performance according to these metrics. YOLOv7, EfficientDet, and EfficientNet also showed significant gains with larger batch sizes, improving the metrics accordingly. VGG16 and ResNet50 work very well on all batch sizes. For example, ResNet50 obtained an accuracy peak of 92.2 % and an F1 score of 92.0 % where the batch size was 32. Progressive improvement is also observed in the Mask R-CNN, which performed best with an accuracy of 94.2 % and an F1 score of 94.0 % for the largest batch. Figure 13 graphically represents these response times for different batch sizes to balance between performance and computational efficiency.

V. CONCLUSION

In person detection for aerial surveillance, various deep learning models is evaluated on accuracy, loss, response time, and occlusion-robustness. YOLOv7 has a general accuracy of 94.0%, with a loss of 0.22 at epoch 30, which is most precise. It provides the best real-time performance with the fastest Frames per Second (FPS), going up to 40 with the lowest latency of 25 ms. The result for YOLOv7 is very impressive, in that it gives an accuracy of 94.0% and a loss of 0.22. It is very efficient in human detectionwith reduced error. Moreover, this model's quick response time makes it most appropriate for real-time surveillance applications. The most remarkable feature of YOLOv7 is maintains an excellent balance among accuracy, low loss, and real-time processing. During this stage, YOLO predicts one bounding box in every location, and for high FPS. This effectiveness is coupled with high efficiency and an accuracy amount without loss, evidenced in the fact by the outstanding YOLOv7 performance metrics. It is very well suited to aerial surveillance systems, where timely and correct detection of a person is important, since it has the capability to provide fine detections with low computational overhead and fast processing speeds. Further improving its suitability to real-world applications in security and monitoring, YOLOv7 has been found to handle high-resolution aerial footage efficiently and be robust against varying environmental conditions.

REFERENCES

1. Maheriya, K., Rahevar, M., Mewada, H., Parmar, M., & Patel, A. (2024). Insights into aerial intelligence: assessing CNN-based algorithms for human action recognition and object detection in diverse environments. *Multimedia Tools and Applications*, 1-43.
2. Sezgin, A., & Boyacı, A. (2024, April). Advancements in Object Detection for Unmanned Aerial Vehicles: Applications, Challenges, and Future Perspectives. In *2024 12th International Symposium on Digital Forensics and Security (ISDFS)* (pp. 1-6). IEEE.
3. Alhafnawi, M., Salameh, H. A. B., Masadeh, A. E., Al-Obiedollah, H., Ayyash, M., El-Khazali, R., & Elgala, H. (2023). A survey of indoor and outdoor uav-based target tracking

Received: August 09 2025

Volume 38 No. 9s , 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

systems: Current status, challenges, technologies, and future directions. *IEEE Access*, *11*, 68324-68339.

4. Al-lQubaydhi, N., Alenezi, A., Alanazi, T., Senyor, A., Alanezi, N., Alotaibi, B., ... & Hariri, S. (2024). Deep learning for unmanned aerial vehicles detection: A review. *Computer Science Review*, *51*, 100614.
5. S. Gokulakrishnan, M A Jarwar, M H Ali, M M Kamruzzaman, I Meenakshisundaram, M M Jaber & R L Kumar (2023). Maliciously roaming person's detection around hospital surface using intelligent cloud-edge based federated learning, *Journal of Combinatorial Optimization*, *45*, no: 13, 2023.
6. Koyun, O. C., Keser, R. K., Akkaya, I. B., & Töreyin, B. U. (2022). Focus-and-Detect: A small object detection framework for aerial images. *Signal Processing: Image Communication*, *104*, 116675.
7. Ahmed, I., Jeon, G., Chehri, A., & Hassan, M. M. (2021). Adapting Gaussian YOLOv3 with transfer learning for overhead view human detection in smart cities and societies. *Sustainable Cities and Society*, *70*, 102908.
8. Azmat, U., Alotaibi, S. S., Al Mudawi, N., Alabduallah, B. I., Alonazi, M., Jalal, A., & Park, J. (2023). An elliptical modeling supported system for human action deep recognition over aerial surveillance. *IEEE Access*, *11*, 75671-75685.
9. Akshatha, K. R., Karunakar, A. K., Shenoy, S., Dhareshwar, C. V., & Johnson, D. G. (2023). Manipal-UAV person detection dataset: A step towards benchmarking dataset and algorithms for small object detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, *195*, 77-89.
10. Bahmanyar, R., & Merkle, N. (2023). Saving lives from above: Person detection in disaster response using deep neural networks. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *10*, 343-350.
11. Khan, M. A., Menouar, H., & Hamila, R. (2023). LCDnet: a lightweight crowd density estimation model for real-time video surveillance. *Journal of Real-Time Image Processing*, *20*(2), 29.
12. Thakur, N., Nagrath, P., Jain, R., Saini, D., Sharma, N., & Hemanth, D. J. (2023). Autonomous pedestrian detection for crowd surveillance using deep learning framework. *Soft Computing*, *27*(14), 9383-9399.
13. Sambolek, S., & Ivasic-Kos, M. (2024). Person Detection and Geolocation Estimation in UAV Aerial Images: An Experimental Approach. In *ICPRAM* (pp. 785-792).
14. Nguyen, H., Nguyen, K., Sridharan, S., & Fookes, C. (2024). AG-ReID. v2: Bridging Aerial and Ground Views for Person Re-identification. *IEEE Transactions on Information Forensics and Security*.
15. Janakiramiah, B., Bethala, S. R., Chereddy, P. C., Ambi, G., & Mohan, C. (2023, June). Building an Intelligent Aerial Surveillance System through Deep-learning based Object Detection. In *2023 8th International Conference on Communication and Electronics Systems (ICCES)* (pp. 1051-1057). IEEE.

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

16. Vasudevan, I., & Nithva, N. S. (2023, July). Automatic Person Detection in Search and Rescue Operations Using Deep Based MultiModel CNN Detectors. In *2023 IEEE World Conference on Applied Intelligence and Computing (AIC)* (pp. 632-637). IEEE.
17. Kim, G. S., Lee, H., Park, S., & Kim, J. (2023). Joint frame rate adaptation and object recognition model selection for stabilized unmanned aerial vehicle surveillance. *ETRI Journal*, *45*(5), 811-821.
18. Zhang, Y., Xu, C., Yang, W., He, G., Yu, H., Yu, L., & Xia, G. S. (2023). Drone-based RGBT tiny person detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, *204*, 61-76.
19. Zhang, H., Shao, F., He, X., Zhang, Z., Cai, Y., & Bi, S. (2023). Research on object detection and recognition method for UAV aerial images based on improved YOLOv5. *Drones*, *7*(6), 402.
20. Sunanda, N., Tumuluru, P., Gayatri, B. N., Rehan, C., Yasaswini, J., & Durga, A. K. (2023, November). Improved Object Detection and Recognition for Aerial Surveillance: Vehicle Detection with Deep Mask R-CNN. In *2023 International Conference on Sustainable Communication Networks and Application (ICSCNA)* (pp. 969-974). IEEE.
21. Akinsemoyin, A., Awolusi, I., Chakraborty, D., Al-Bayati, A. J., & Akanmu, A. (2023). Unmanned aerial systems and deep learning for safety and health activity monitoring on construction sites. *Sensors*, *23*(15), 6690.
22. Aibibu, T., Lan, J., Zeng, Y., Lu, W., & Gu, N. (2023). An efficient rep-style gaussian–wasserstein network: Improved uav infrared small object detection for urban road surveillance and safety. *Remote Sensing*, *16*(1), 25.
23. Zhang, X., Feng, Y., Zhang, S., Wang, N., Mei, S., & He, M. (2023). Semi-supervised person detection in aerial images with instance segmentation and maximum mean discrepancy distance. *Remote Sensing*, *15*(11), 2928.
24. Hanzla, M., Yusuf, M. O., & Jalal, A. (2024, May). Vehicle Surveillance Using U-NET Segmentation and DeepSORT Over Aerial Images. In *2024 International Conference on Engineering & Computing Technologies (ICECT)* (pp. 1-6). IEEE.
25. Stan, A. S., Ichim, L., Parvu, V. P., & Popescu, D. (2023, June). Person Detection and Tracking Using UAV and Neural Networks. In *2023 31st Mediterranean Conference on Control and Automation (MED)* (pp. 323-328). IEEE.
26. Valarmathi, B., Kshitij, J., Dimple, R., Srinivasa Gupta, N., Harold Robinson, Y., Arulkumaran, G., & Mulu, T. (2023). Human detection and action recognition for search and rescue in disasters using yolov3 algorithm. *Journal of Electrical and Computer Engineering*, *2023*(1), 5419384.
27. Zhang, Y., Yin, Y., & Shao, Z. (2023). An Enhanced Target Detection Algorithm for Maritime Search and Rescue Based on Aerial Images. *Remote Sensing*, *15*(19), 4818.
28. Patel, N., Vasani, N., Gupta, R., Kumar Jadav, N., Tanwar, S., Alqahtani, F., ... & Simona Raboaca, M. (2023). Convolutional neural network and unmanned aerial vehicle-based public safety framework for human life protection. *International Journal of Communication Systems*, e5545.

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

29. Kucukayan, G., & Karacan, H. (2024). YOLO-IHD: Improved Real-Time Human Detection System for Indoor Drones. *Sensors*, *24*(3), 922.
30. Abood, S. S., Hussein, K. Q., & Gaata, M. T. (2023). Design a border Surveillance System based on Autonomous Unmanned Aerial Vehicles (UAV). *Al-Iraqia Journal for Scientific Engineering Research*, *2*(4), 51-62.
31. Xu, S., Luo, L., Hong, H., Hu, J., Yang, B., & Hu, S. (2024). Multi-granularity attention in attention for person re-identification in aerial images. *The Visual Computer*, *40*(6), 4149-4166.
32. Rahman, S., Rony, J. H., Uddin, J., & Samad, M. A. (2023). Real-Time Obstacle Detection with YOLOv8 in a WSN Using UAV Aerial Photography. *Journal of Imaging*, *9*(10), 216.
33. Khel, M. H. K., Kadir, K. A., Khan, S., Noor, M., Nasir, H., Waqas, N., & Khan, A. (2023). Realtime Crowd Monitoring—Estimating Count, Speed and Direction of People Using Hybridized YOLOv4. *IEEE Access*, *11*, 56368-56379.
34. Wang, S., Jiang, H., Li, Z., Yang, J., Ma, X., Chen, J., & Tang, X. (2024). PHSI-RTDETR: A Lightweight Infrared Small Target Detection Algorithm Based on UAV Aerial Photography. *Drones*, *8*(6), 240.
35. Cheng, P., Xiong, Z., Bao, Y., Zhuang, P., Zhang, Y., Blasch, E., & Chen, G. (2023). A deep learning-enhanced multi-modal sensing platform for robust human object detection and tracking in challenging environments. *Electronics*, *12*(16), 3423.
36. Hong, C. J., Aparow, V. R., & Jamaluddin, H. (2023). Real-time human search and monitoring system using unmanned aerial vehicle. *International Journal of Vehicle Autonomous Systems*, *17*(1-2), 106-132.
37. Pathak, S., Shrestha, S., & AlMahmoud, A. (2024). Model Agnostic Defense against Adversarial Patch Attacks on Object Detection in Unmanned Aerial Vehicles. *arXiv preprint arXiv:2405.19179*.
38. Tang, G., Ni, J., Zhao, Y., Gu, Y., & Cao, W. (2023). A survey of object detection for UAVs based on deep learning. *Remote Sensing*, *16*(1), 149.
39. Ouardirhi, Z., Mahmoudi, S. A., & Zbakh, M. (2024). Enhancing Object Detection in Smart Video Surveillance: A Survey of Occlusion-Handling Approaches. *Electronics*, *13*(3), 541.
40. Sary, I. P., Andromeda, S., & Armin, E. U. (2023). Performance Comparison of YOLOv5 and YOLOv8 Architectures in Human Detection using Aerial Images. *Ultima Computing: Jurnal Sistem Komputer*, *15*(1), 8-13.
41. Bouhlel, F., Mliki, H., & Hammami, M. (2023, August). Person Activity Classification from an Aerial Sensor Based on a Multi-level Deep Features. In *International Conference on Advanced Concepts for Intelligent Vision Systems* (pp. 66-75). Cham: Springer Nature Switzerland.
42. Murugan, R. A., & Sathyabama, B. (2023). Object detection for night surveillance using Ssan dataset based modified Yolo algorithm in wireless communication. *Wireless Personal Communications*, *128*(3), 1813-1826.
43. Akhtar, I., Al Mudawi, N., Alabdullah, B. I., Alonazi, M., & Park, J. (2023). Human-based Interaction Analysis via Automated Key point Detection and Neural Network Model. *IEEE Access*.

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

44. Singh, A., Prakash, K., Manda, P. A., Sona, D. R., & Das, R. R. (2024). Development of an Autonomous UAS for on Air Surveillance and Object Detection: A Real Execution. *Journal of Electrical Engineering & Technology*, *19*(1), 723-737.
45. Othman, N. A., & Aydin, I. (2023). Development of a novel lightweight CNN model for classification of human actions in UAV-captured videos. *Drones*, *7*(3), 148.
46. Olaniyi, O., & Ganiyu, S. (2023). Intelligent Video Surveillance System Using Faster Regional Convolutional Neural Networks. *Balkan Journal of Electrical and Computer Engineering*, *11*(4), 346-351.
47. Aibibu, T., Lan, J., Zeng, Y., Lu, W., & Gu, N. (2024). Feature-Enhanced Attention and Dual-GELAN Net (FEADG-Net) for UAV Infrared Small Object Detection in Traffic Surveillance. *Drones*, *8*(7), 304.
48. Onifade, T. O. (2023). *An Optimized Object Detection Model for Unmanned Aerial Vehicles*. University of Louisiana at Lafayette.
49. Saeed, S. M., Akbar, H., Nawaz, T., Elahi, H., & Khan, U. S. (2023). Body-Pose-Guided Action Recognition with Convolutional Long Short-Term Memory (LSTM) in Aerial Videos. *Applied Sciences*, *13*(16), 9384.
50. Madulara, R. F., Aleluya, E. R. M., Pao, J. C., Clar, S. E., Alagon, F. J. A., Salaan, C. J. O., & Bahinting, M. F. P. (2023, November). Detecting Persons in River Quarry Operations: Aerial Surveillance for Flood Pre-Disaster Monitoring. In *2023 IEEE 15th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)* (pp. 1-6). IEEE.
51. Dilshad, N., Khan, T., & Song, J. (2023). Efficient Deep Learning Framework for Fire Detection in Complex Surveillance Environment. *Comput. Syst. Sci. Eng.*, *46*(1), 749-764.
52. Zhang, Y., Tao, Q., & Yin, Y. (2023). A Lightweight Man-Overboard Detection and Tracking Model Using Aerial Images for Maritime Search and Rescue. *Remote Sensing*, *16*(1), 165.
53. Xu, L., Yang, Q., Qin, M., Wu, W., & Kwak, K. (2023). Collaborative Human Recognition with Lightweight Models in Drone-based Search and Rescue Operations. *IEEE Transactions on Vehicular Technology*.
54. Zhou, W., Cai, C., Zheng, L., Li, C., & Zeng, D. (2024). ASSD-YOLO: a small object detection method based on improved YOLOv7 for airport surface surveillance. *Multimedia Tools and Applications*, *83*(18), 55527-55548.
55. Han, T., Dong, Q., & Sun, L. (2023). SenseLite: A YOLO-Based Lightweight Model for Small Object Detection in Aerial Imagery. *Sensors*, *23*(19), 8118.
56. Zhao, D., Shao, F., Yang, L., Luo, X., Liu, Q., Zhang, H., & Zhang, Z. (2023). Object Detection Based on an Improved YOLOv7 Model for Unmanned Aerial-Vehicle Patrol Tasks in Controlled Areas. *Electronics*, *12*(23), 4887.
57. Sinha, K. P., & Kumar, P. (2023). Human activity recognition from UAV videos using a novel DMLC-CNN model. *Image and Vision Computing*, *134*, 104674.
58. Xu, L., Peng, H., Lu, X., & Xia, D. (2023). Learning to generalize aerial person re-identification using the meta-transfer method. *Concurrency and Computation: Practice and Experience*, *35*(12), e7687.

Volume 38 No. 9s , 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

59. Daramouskas, I., Meimetis, D., Patrinopoulou, N., Lappas, V., Kostopoulos, V., & Kapoulas, V. (2023). Camera-based local and global target detection, tracking, and localization techniques for UAVs. *Machines*, *11*(2), 315.
60. Liang, B., Su, J., Feng, K., Liu, Y., & Hou, W. (2023). Cross-layer triple-branch parallel fusion network for small object detection in uav images. *IEEE Access*, *11*, 39738-39750.
61. Dataset Collection:Kaggle repository- https://www.kaggle.com/datasets/constantinwerner/human-detection-dataset
62. Putra, M. H., Yussof, Z. M., Salim, S. I., & Lim, K. C. (2017). Convolutional neural network for person detection using yolo framework. Journal of Telecommunication, Electronic and Computer Engineering (JTEC), 9(2-13), 1-5.
63. Wei, H., & Kehtarnavaz, N. (2019). Semi-supervised faster RCNN-based person detection and load classification for far field video surveillance. Machine Learning and Knowledge Extraction, 1(3), 44.
64. Kim, C. E., Oghaz, M. M. D., Fajtl, J., Argyriou, V., & Remagnino, P. (2018). A comparison of embedded deep learning methods for person detection. arXiv preprint arXiv:1812.03451.
65. Luo, C. Y., Cheng, S. Y., Xu, H., & Li, P. (2022). Human behavior recognition model based on improved EfficientNet. Procedia computer science, 199, 369-376.
66. Minkesh, A., Worranitta, K., & Taizo, M. (2019). Human extraction and scene transition utilizing Mask R-CNN. arXiv preprint arXiv:1907.08884.

Received: August 09 2025