International Research Journal of Multidisciplinary Scope (IRJMS), 2026; 7(1): 1040-1049 IRJMS

Original Article | ISSN (O): 2582-631X DOI: 10.47857/irjms.2026.v07i01.07685

Multi Object Detection System for Video Surveillance Using an Improved Yolov5 Model

Ahila Maheswari M[1]*, Rajesh S[1], Jeyapandi Marimuthu[2]

[1]Department of Information Technology, Mepco Schlenk Engineering College, Sivakasi, India, [2]Sethu Institute of Technology, Kariyapatti, India. *Corresponding Author's Email: ahilakshmi@gmail.com

Abstract

Rapid advancements in artificial intelligence have made video surveillance increasingly pervasive across public and private domains. Multi-object detection (MOD) has emerged as a key research focus within video surveillance due to its critical role in security monitoring, crowd analysis, and anomaly detection. Traditional MOD systems rely on machine learning-based pipelines that typically follow a divide-and-conquer strategy for parameter optimization. However, these methods often exhibit limited performance due to constraints in model architecture and feature representation. To address these challenges, this paper proposes an enhanced You Only Look Once Version 5 (YOLOv5) framework, termed Attention-based YOLOv5 (AYOLOv5), specifically optimized for MOD in surveillance videos. The proposed system integrates attention mechanisms to refine feature extraction, improving both the detection accuracy and computational efficiency. Initially, the surveillance video frames undergo pre-processing steps, including frame conversion and data augmentation, to enrich the dataset and improve model generalization. Subsequently, the AYOLOv5 model detects multiple objects, leveraging a fuzzy c-means (FCM) clustering approach to optimize anchor box generation. Experimental evaluations conducted on the MOT20 dataset demonstrate that the proposed framework achieves a superior detection accuracy of 98.90%, outperforming existing state-of-the-art models. These results highlight the model's effectiveness in handling complex scenarios involving occlusions, overlapping objects, and varying object scales, thereby significantly enhancing the reliability and practical utility of video surveillance systems for real-world applications.

Keywords: Multi-object Detection, Transfer Learning, Video Surveillance, You Look Only Once.

Introduction

Multiple Object Detection (MOD), using computer vision plays a crucial role in analysing video streams by identifying and localizing multiple objects over time. MOD has become a core component of modern applications such as intelligent surveillance systems, traffic monitoring, autonomous driving, and medical image analysis (1, 2). Despite notable progress, accurate detection of multiple objects remains challenging due to occlusions, scale variations, appearance changes, and interactions among objects in crowded scenes (3). The widespread deployment of CCTV and large-scale surveillance infrastructures has significantly increased the demand for automated, real-time detection systems. Manual monitoring of video feeds is inefficient and prone to human error, especially in control centers operating continuously (4). Consequently, robust and real-time MOD frameworks are essential for effective situational awareness and decision-making. Recent advances in deep learning (DL) have greatly improved object detection performance. YOLOv4 further improved speed and accuracy by integrating CSPDarknet53 as the backbone, along with data augmentation and optimization strategies (5). Traditional DL-based detectors often involve multi-stage pipelines and repeated feature extraction at different scales, leading to increased computational complexity and latency. Moreover, these models tend to focus on local regions of interest, limiting their ability to capture global contextual information (6). To overcome these limitations, single-stage detectors such as You Only Look Once (YOLO) have gained significant attention. YOLO reformulates object detection as a regression problem, enabling simultaneous prediction of bounding boxes and class probabilities in a single forward pass. This design enables real-time performance while maintaining competitive accuracy (7). Among the

(Received 19th August 2025; Accepted 02nd January 2026; Published 28th January 2026)

YOLO family, YOLOv5 stands out due to its improved training efficiency, lightweight architecture, and superior detection accuracy, making it suitable for real-time MOD applications in dynamic environments (8). Nevertheless, YOLOv5 still faces challenges in complex surveillance scenarios involving overlapping objects, background clutter, and partial occlusions. Attention mechanisms have recently emerged as an effective solution to enhance feature representation by emphasizing relevant spatial regions while suppressing irrelevant information (9). Additionally, accurate bounding box localization remains critical for improving MOD performance. Integrating clustering-based approaches, such as Fuzzy C-Means (FCM), can improve bounding box refinement by better handling ambiguity in object boundaries, particularly in crowded scenes (10).

Motivated by these observations, this work enhances the YOLOv5 architecture by incorporating a soft attention mechanism within the backbone using an inception-based design, along with an FCM-based clustering strategy for bounding box detection. These enhancements aim to improve detection accuracy and robustness in challenging surveillance video scenarios.

This research focuses on improving the effectiveness of YOLOv5 for multi-object detection in real-world surveillance scenarios. The study aims to enhance the model's ability to capture meaningful visual features by embedding a soft attention mechanism into the backbone network, allowing the detector to prioritize important regions within complex scenes. To further refine detection performance, an FCM-based clustering strategy is introduced to achieve more precise bounding box localization, especially in crowded environments where objects often overlap. The proposed approach is thoroughly evaluated in challenging surveillance conditions involving occlusions and dense object interactions, highlighting its robustness and practical applicability for intelligent video surveillance systems.

The remainder of this paper is organized as follows: Section 2 reviews related works in MOD and YOLO-based detection. Section 3 details the proposed methodology. Section 4 presents experimental results and comparative analysis, and Section 5 concludes the study with future research directions.

Several studies have investigated multi-object detection (MOD) and tracking in surveillance videos using deep learning techniques. A hybrid framework combining LuNet for feature extraction and YOLOv2 for object detection was introduced, achieving an accuracy of 94% on the MOT20 dataset (11). An enhanced YOLOv8-based pedestrian tracking system was developed by integrating soft Non-Maximum Suppression, GhostConv, and C3Ghost modules, and it was evaluated on the MOT17 and MOT20 datasets, yielding accuracies of 61.077% and 48.326%, respectively (12). Improvements to the YOLOv5-7S architecture were proposed for high-resolution video processing by employing multi-GPU configurations and adding specialized layers to enhance the detection of small objects (13). In another approach, faster R-CNN combined with an Inception-v2 backbone and a unifocal KLT tracking method was utilized to track multiple objects in densely populated video sequences (14).

Furthermore, a modified YOLOv5 architecture was applied for pedestrian detection to enhance detection performance in surveillance environments (15). High-resolution feature extraction using HRNet integrated with polarized self-attention and circle loss was employed to improve discriminative feature learning for re-identification tasks (16). A robust MODT framework was proposed by combining a path-augmented RetinaNet with quasi-recurrent neural networks, enabling accurate object organization in complex scenes (17). Occluded object detection was addressed using motion track prediction and spatio-temporal feature modeling, demonstrating superior performance on the MOT16, MOT17, and MOT20 datasets (18). Additionally, several encoder–decoder and CNN-based architectures were proposed to improve detection accuracy and robustness in surveillance videos (19–21). Multi-scale traffic object detection was enhanced by integrating SPD-Conv and normalized attention mechanisms into the YOLOv5 framework, resulting in a 7.1% improvement in accuracy (22). A lightweight convolutional encoder–decoder segmentation network was also introduced to reduce network parameters while improving detection performance on the CDNet 2012 dataset (23).

Research Gaps

Even though the existing models use recent approaches, they still have several limitations. Some work focuses on something other than data augmentation. The DL model depends on higher volumes of diverse data to establish accurate detection in many contexts. Data augmentation increases the generation of data variation, which helps the model to improve detection accuracy. Later, some existing models will use the YOLO method, which uses a CNN technique as a feature extractor to forecast the object's bounding boxes and class probabilities in the input data. However, a considerable amount of data is required for CNN to extract the beneficial features exclusive to computational and memory-intensive training. The pre-trained CNN models such as VGG19 and LuNet models cannot capture delicate-grained contextual information within surveillance videos, causing sub-optimal performance in object detection. Also, the techniques face challenges in localizing the objects within the image accurately for scenarios like occlusions and overlapping objects. To overcome the existing limitations, our study develops an attention-based YOLOV5 network for improving the accuracy and robustness of MOD in surveillance videos and an FCM-based clustering technique for object localization in YOLOV5, which makes the system quicker and more effective. A detailed explanation of the proposed system is given in the sections below.

Methodology

Figure 1 illustrates the proposed system's workflow for MOD in surveillance videos. It primarily includes two stages: pre-processing, which contains frame conversion and data augmentation, and MOD, which are described in the following subsection.

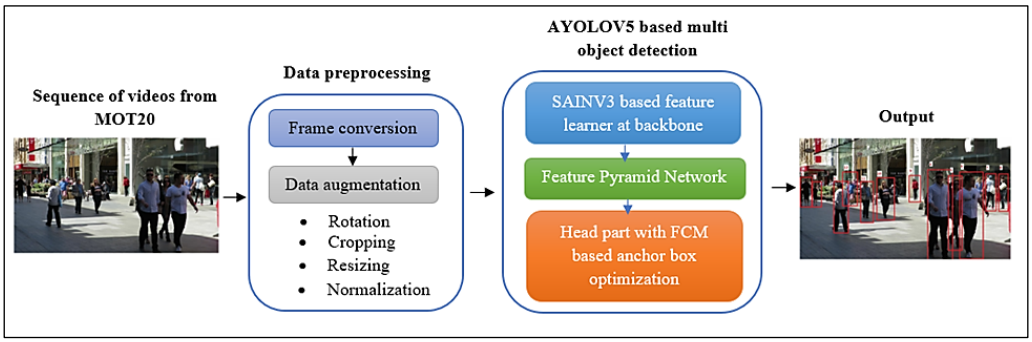

Figure 1: Workflow of the Proposed Methodology

Preprocessing

Firstly, the input data is taken from the openly available dataset. Due to the impact of climate and lighting environment, video during transmission can distort the drop-in camera capture video quality. Pre-processing is an essential task in video processing, and it processes the original data to detect multi-objects accurately. We are performing two types of pre-processing procedures, such as frame conversion and data augmentation, to avoid class imbalance and over fitting issues.

Frame Conversion

The surveillance videos are in several formats and resolutions. Converting the videos into frames standardizes the video's resolution, data format, and frame rate and enables consistency across the dataset. It also reduces file size and storage needs, which helps with faster data processing during training, particularly for large video datasets.

Data Augmentation

A considerable quantity of training data is required to avoid over fitting and enhance the model's performance. Data augmentation enhances the training data required to implement the learning model for MOD. Frequently used data augmentation techniques, including rotation, cropping, resizing, and normalization, are used in our current study, which increases diversity. Hence, the model generalizes well on unseen scenarios, improving the robustness to varying illumination conditions, object orientations, and background clutter. It also balances the dataset by generating augmented samples for underrepresented classes, which gives the model enough data for each class label of the dataset, causing a balanced performance among several object classes. The steps we considered for augmentation are as follows:

Rotation: The augmented image can be improved by using random rotation. This means that the image can be rotated at any direction of 90 degrees for better flexibility.

Cropping: It removes the image's irrelevant or noisy regions. It's frequently used in object detection, where the area of interest wants to be isolated.

Resizing: The input image dimensions are set to 299x299x3. The foremost reason for selecting this dimension is to preserve the balance among computational efficiency and image details for detection.

Normalization: It scales the image's pixel values to a specific range of 0 to 1. It ensures that the values are constant, which makes the model easy to learn and decreases the effect of changing illumination situations. The min-max normalization is utilized to normalize the data. It is formulated as [1]:

$$\overrightarrow{V}_{NR} = \frac{\dot{N} - min(\overrightarrow{V})}{max(\overrightarrow{V}) - min(\overrightarrow{V})} \quad [1]$$

Here, $\overrightarrow{V}_{NR}$ refers to the normalization factor, $\overrightarrow{V}$ signifies the numerically selected value, $min(\overrightarrow{V})$ and $max(\overrightarrow{V})$ represents the minimum as well as the maximum value of the feature, correspondingly.

Multi-Object Detection

After preprocessing, MOD is done using AYOLOV5. YOLOv5 is a widely recognized MOD framework for its performance and speed. It has a vibrant and flexible structure that can be broken down, adjusted, and made on a commonly available platform. YOLOV5 suggests a lightweight object detection procedure for robots with enhanced speed and less computation time than the other models. In this network, the model contains three parts. The backbone network (BBN), the feature pyramid network (FPN), and the detection head network (DHN). The BBN is used to extract features from the images at various scales. Second, the FPN combines the various scale features and transfers them to the detection network. Third, the DHN is used to predict the object category that uses the features of the image and makes the object bounding box. The existing YOLOV5 model's performance is enhanced by using the soft attention-based InceptionV3 (SAINV3) model as a feature extractor at the backbone part. This expands the ability of the network to capture fine-grained details and contextual information of the input data at different scales by replacing the default Dark net model.

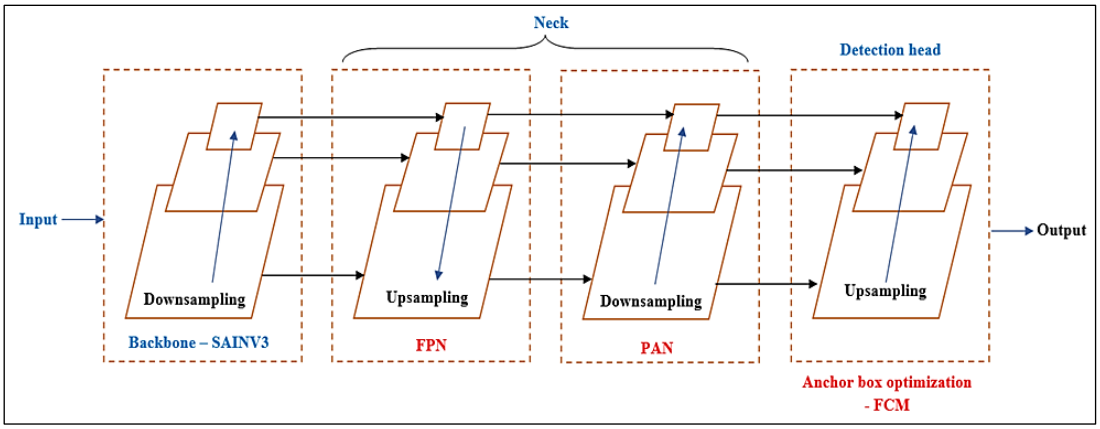

Figure 2: Structure of the YOLOV5

Second, the FCM clustering procedure establishes the network's anchor size to simplify the bounding-box regression while decreasing model parameters and computational load. These improvements over traditional YOLOV5 networks enhance the model's performance for MOD.

Figure 2 mostly comprises three parts: backbone, neck, and head parts. These parts are briefly explained as follows:

Backbone Part

It is the element that takes the input image and extracts its feature maps. In this part, the SAINV3 model is used instead of the default model for feature learning. InceptionV3 (INV3) is the improved version of the inception family that made some enhancements with factorized 7x7 convolutions, label smoothing, and the auxiliary classification usage for spreading label data less to

the network. The batch normalizations are mainly used for image analysis, object detection, and layers in the side head. Still, the conventional INV3 does not capture relevant contextual information required for MOD. Hence, for accurate object detection, soft attention is included in INV3 to capture relevant contextual and fine-grained information from the input data at different relevant scales. Additionally, with an attention mechanism, the feature extractor is beneficial in extracting the input signal's features by manipulating inter-channel and inter-temporal-spatial relationships and highlighting vital, deeper information. The proposed SAINV3 process is given as follows:

The INV3's initial layer contains 3 standard convolution layers with 3x3kernel size. This convolutional layer convolves the input image (299x299x3) with a set of learnable parameters to encode certain properties into the architecture. Then, the convoluted features are transferred to the max-pooling layer with a 3x3 kernel size. This layer down samples the feature map recollects the non-trivial features, and captures the maximum value within the convoluted feature map. Then, these pooled feature maps are transferred to the two 3x3 convolutional layers for more extracting features. The network's next stage contains an inception convolution that concurrently convolutes the features using various filter sizes for every convolution, combining or stacking the outcomes together and passing them across the network. In INV3, 3 Inception A modules, 5 Inception B modules, and 2 Inception C modules are collected in sequence.

After the Inception modules and the convolutional layers, the feature maps are transferred to the SA layer to extract the deep hidden features. The soft attention module obtains the features and enhances the feature output by removing the deeper features (24). The attention module delivers a weight for features, which can be contributed in back propagation. It is calculated by taking the matrix multiplication among attention weights with a feature vector is achieved. The attention vector $(\ddot{S}\ddot{A}_S)$is computed to extract the deeper features by using equation [2].

$$\ddot{S}\ddot{A}_S = \sum_{k=1}^{N} \beta^{\leftarrow}{}_k \ddot{P}_k \quad [2]$$

Where, $\beta^{\leftarrow}{}_k$indicates a nonlinear SoftMax function and $\ddot{P}_k$refers to the extracted features of the previous layer. Finally, the feature map is flattened into dimensions were 5×5 with 2,048 channels.

Neck Part

This part is between the backbone and the prediction head and is primarily used to combine the image features extracted by the backbone. The FPN layer transfers from top to bottom on a string semantic feature. It mainly connects the upper- and lower-layer features to enhance the MOD performance, specifically for multi-objects. The backbone and neck's information are linked via a horizontal connection, which preserves more information and allows communication among the feature's adjacent levels. The PAN successfully spreads healthy localization features from the bottom to the top layers. It integrates the parameters from several detection and backbone layers to preserve more low-level features.

Head Part

This part is responsible for captivating feature maps and inferring the bounding boxes and classes by taking in numerous aggregated feature maps from the neck. The anchor box enables a network to detect multiple objects, several scale objects, and overlapping objects. The sizes of the anchor are selected manually in YOLOV5. Any size of anchor boxes can be generated by the network on the bounding-box regression to estimate a neighbouring ground-truth box and better detections are easy to learn and predict only if we select improved anchor sizes. In the bounding boxes training set, the FCM process is used to decide the anchor size automatically rather than selecting manually, and it is a data clustering procedure where every data point belongs to the definite degree's cluster by a membership grade (25). The clustering technique permits the model to choose optimal anchor box sizes based on the dataset's object shapes and size distribution. This ensures that the anchor boxes cover a broad range of object scales and aspect ratios, improving the model's power to detect objects of various sizes more effectively.

Experimental Setup

The proposed model was evaluated against existing object detection approaches using standard evaluation metrics, namely accuracy, precision, recall, and F1-score (26). All experiments were conducted on a system configured with Ubuntu 21.04, PyTorch 1.8.0, Python 3.8, an 11th Gen Intel® Core™ i5-11400

CPU operating at 2.60 GHz, NVIDIA GeForce RTX 3070 GPU, CUDA 11.2, cuDNN 7.6, and 16 GB RAM. This configuration ensured sufficient computational capability for training and testing deep learning-based MOD models under real-time constraints.

Dataset Description and Reliability

The performance of the proposed model was evaluated using the publicly available MOT20 dataset, which is widely recognized as a benchmark dataset for multi-object detection and tracking tasks in crowded surveillance scenarios. The dataset was obtained from the official MOT Challenge repository, ensuring authenticity and reproducibility (https://motchallenge.net/data/MOT20/).

MOT20 consists of eight high-density pedestrian video sequences captured under unconstrained real-world conditions, including severe occlusions, illumination variations, and diverse viewpoints. Four sequences are designated for training and four for testing. The testing sequences include crowded square scenes at night time (CSNT), pedestrian exits from stadium environments (PEN), crowded indoor train stations (CTS), and pedestrian street scenes (PSS). The dataset is manually annotated with high-quality bounding boxes and has been extensively validated in prior studies, making it a reliable and standardized benchmark for evaluating MOD performance in surveillance applications.

Qualitative Analysis

Figure 3 illustrates sample detection outputs of the proposed model on the MOT20 dataset. The detected frames demonstrate accurate localization and classification of multiple objects under challenging conditions such as heavy crowd density and partial occlusions, highlighting the robustness of the proposed approach.

Figure 3: (A) Original Frames, (B) Object Detected Frames, Detection Results of the Proposed System

Results

Quantitative Performance Evaluation

This section presents a comparative analysis of the proposed model against YOLOv5, Faster R-CNN (F-RCNN), InceptionV3, and CNN-based approaches across CSNT, PEN, CTS, and PSS video sequences (27-28).

Figure 4 illustrates the accuracy comparison. The CNN-based approach exhibits relatively lower accuracy (91.56% and 91.34%) due to its limited ability to capture multi-scale contextual features and its dependency on large annotated datasets. Although YOLOv5, F-RCNN, and InceptionV3 achieve competitive results, their reliance on raw feature representations limits fine-grained contextual understanding.

In contrast, the proposed AYOLOv5 model achieves superior accuracy values of 98.95%, 98.74%, 99.06%, and 98.84% for CSNT, PEN, CTS, and PSS sequences, respectively, demonstrating its effectiveness in multi-object detection under complex surveillance conditions. Table 1 presents a detailed comparison of precision, recall, and F1-score. Across all sequences, the proposed model consistently outperforms the existing methods, achieving the highest values in all evaluation metrics.

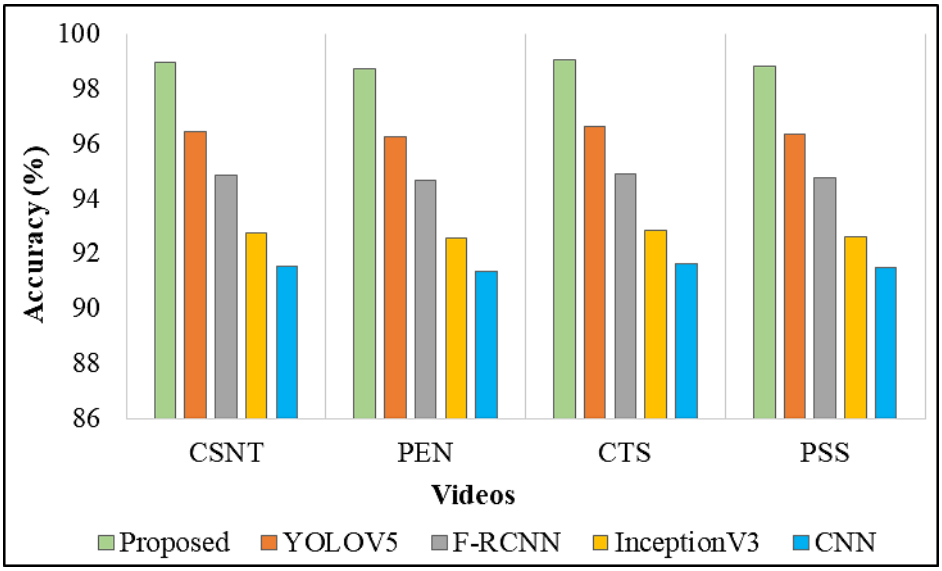

Figure 4: Accuracy Analysis of the Proposed and Existing Methods

Table 1: Results of Classifiers Regarding Precision, Recall and F-Score

Metrics	Video	Proposed	YOLOV5	F-RCNN	InceptionV3	CNN
Precision	CSNT	99.06	96.52	94.92	92.85	91.64
	PEN	98.82	96.31	94.74	92.64	91.42
	CTS	99.12	96.72	94.98	92.93	91.73
	PSS	98.91	96.42	94.82	92.69	91.52
Recall	CSNT	98.86	96.39	94.79	92.71	91.47
	PEN	98.67	96.17	94.57	92.48	91.26
	CTS	98.96	96.58	94.85	92.78	91.58
	PSS	98.78	96.28	94.68	92.52	91.36
F1-score	CSNT	98.98	96.71	94.89	92.82	91.59
	PEN	98.78	96.28	94.71	92.59	91.38
	CTS	99.09	96.68	94.96	92.89	91.69
	PSS	98.87	96.39	94.79	92.65	91.52

Discussion

Comparative Analysis with Existing Studies

The obtained results align with recent research trends emphasizing attention mechanisms and contextual feature learning for improved MOD performance. Compared to traditional YOLOv5-based approaches reported in earlier studies, the proposed method demonstrates a noticeable improvement in precision and recall, particularly in highly crowded scenes such as CTS and CSNT. This improvement is consistent with findings that highlight the effectiveness of attention-based architectures in suppressing background noise and enhancing salient object features. While Faster R-CNN and Inception-based models have shown robustness in controlled environments, their performance degradation in dense and occluded scenes observed in this study corroborates earlier findings that multi-stage detectors struggle with real-time scalability and contextual ambiguity. The superior performance of the proposed model can be attributed to the integration of a soft attention-based inception backbone, which enables multi-scale feature extraction while preserving global contextual information. Additionally, the use of FCM-based clustering for bounding box optimization addresses localization ambiguities commonly reported in earlier YOLO-based detectors, particularly under overlapping object scenarios. This explains the observed improvement over existing YOLO variants that rely on predefined anchor box configurations.

Ethical Considerations

The proposed research strictly adheres to ethical standards associated with computer vision and surveillance-based systems. The MOT20 dataset used in this study is publicly available, anonymized, and collected for academic research

purposes, ensuring compliance with data protection and privacy regulations. No personally identifiable information (PII) is processed or inferred at any stage of the experimentation. Furthermore, the proposed system is intended to assist automated monitoring tasks rather than replace human judgment, thereby reducing risks associated with biased or autonomous decision-making. The design avoids facial recognition or identity inference, focusing solely on object-level detection, which minimizes ethical concerns related to individual profiling and privacy invasion. These considerations ensure responsible and ethical deployment in real-world surveillance applications.

Practical Implications and Recommendations

The findings of this study have significant practical implications for real-time surveillance systems deployed in high-density environments such as transportation hubs, public events, and smart cities. The improved detection accuracy and robustness under occlusions make the proposed model suitable for enhancing public safety monitoring and crowd management systems.

It is recommended that future implementations integrate the proposed framework with edge-computing devices to enable decentralized and scalable surveillance solutions. Additionally, extending the model to support multi-object tracking and behaviour analysis could further enhance its applicability in intelligent security systems. Future research may also explore lightweight attention mechanisms to reduce computational overhead, facilitating deployment on resource-constrained devices.

Conclusion

This paper presented **AYOLOv5**, an enhanced deep learning framework for multi-object detection (MOD) in surveillance videos. The proposed approach was designed to address critical challenges in dense and dynamic environments, including occlusions, overlapping objects, and scale variations. The effectiveness of the model was validated using the publicly available MOT20 benchmark dataset, which represents highly crowded real-world surveillance scenarios. Comprehensive experimental evaluations were conducted by comparing the proposed method with existing models, including YOLOv5, Faster R-CNN, InceptionV3, and conventional CNN-based approaches, using standard performance metrics such as accuracy, precision, recall, and F1-score.

The primary contribution of this work lies in the integration of a soft attention-based inception backbone within the YOLOv5 architecture, enabling enhanced multi-scale feature representation and improved contextual awareness. Additionally, the incorporation of an FCM-based clustering strategy for bounding box optimization significantly improved object localization, particularly in challenging scenes with heavy occlusions and object overlap. Experimental results demonstrated that the proposed AYOLOv5 model consistently outperformed the compared methods across all evaluated video sequences, namely CSNT, PEN, CTS, and PSS. Notably, superior performance was achieved in the CTS sequence, with accuracy, precision, recall, and F1-score reaching 99.06%, 99.12%, 98.96%, and 99.09%, respectively, highlighting the robustness of the proposed approach in densely populated environments.

From a practical perspective, the findings of this study indicate that AYOLOv5 can serve as a reliable solution for real-time surveillance applications, including crowd monitoring, public safety management, and intelligent transportation systems. The improved detection accuracy and robustness under complex conditions make the model suitable for deployment in large-scale surveillance infrastructures where precise and timely object detection is critical.

Despite its promising performance, the proposed approach has certain limitations. The model was evaluated primarily on a single benchmark dataset focusing on pedestrian-dense scenes, which may limit its generalizability to other object categories or less structured environments. Furthermore, the integration of attention mechanisms and clustering techniques introduces additional computational overhead, which may impact deployment on resource-constrained edge devices. Future research will focus on extending the proposed framework to larger and more diverse datasets to enhance generalization capability. Incorporating lightweight attention mechanisms and model compression techniques can further reduce computational complexity, enabling deployment on embedded and edge-based platforms. Additionally, integrating multi-object

tracking and behaviour analysis modules may broaden the applicability of the proposed system to advanced intelligent surveillance and smart city applications.

Abbreviations

BBN: backbone network, CNN: convolutional neural networks, CSNT: crowded square by night time, CTS: crowded indoor train station, DHN: Detection head network, FCM: fuzzy c-means, FPN: feature pyramid network, PEN: people who leave the stadium's entrance by night time, PSS: pedestrian street scene, RNN: recurrent neural networks, SAINV3: soft attention-based InceptionV3, YOLOV5: You Look Only Once Version 5.

Acknowledgement

None.

Author Contributions

Ahila Maheswari M: analysis, conceptualization, model design, data collection, formulation, testing, Rajesh S: analysis, data collection, testing, Jeyapandi Marimuthu: analysis, testing, validation.

Conflict of Interest

The authors declare that there are no conflicts of interest related to this research work. No financial, personal, or professional relationships have influenced the findings, analysis, or conclusions presented in this study.

Declaration of Artificial Intelligence (AI) Assistance

The authors confirm that no generative artificial intelligence or AI-assisted tools were used in the preparation or writing of this manuscript. All content was created by the authors.

Ethics Approval

This study was conducted in accordance with the ethical guidelines and principles.

Funding

None.

References

1. Jiao L, Zhang F, Liu F, *et al*. A survey of deep learning-based object detection. IEEE Access. 2019; 7: 12883 7–128868. doi:10.1109/ACCESS.2019.2939201
2. Liu L, Ouyang W, Wang X, Fieguth P, Chen J, Liu X, Pietikäinen M. Deep learning for generic object detection: a survey. Int J Comput Vis. 2020;128(2):261–318. doi:10.1007/s11263-019-01247-4
3. Wang CY, Bochkovskiy A, Liao HYM. Scaled-YOLOv4: Scaling Cross Stage Partial Network. Proc IEEE/CVF Conf Comput Vis Pattern Recognit. 2021:13029–13038. doi:10.1109/CVPR46437.2021.01283
4. Sreenu G, Saleem Durai MA. Intelligent video surveillance: a review through deep learning techniques for crowd analysis. J Big Data. 2019; 6:48. doi:10.1186/s40537-019-0212-5
5. Redmon J, Farhadi A. YOLOv3: An incremental improvement. arXiv preprint arXiv:1804.02767. 2018. https://arxiv.org/pdf/1804.02767
6. Bochkovskiy A, Wang CY, Liao HYM. YOLOv4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934. 2020. https://arxiv.org/pdf/2004.10934
7. Jocher G, Stoken A, Borovec J, *et al*. YOLOv5 by Ultralytics. Zenodo. 2022. doi: 10.5281/zenodo.6222936
8. Olorunshola OE, Irhebhude ME, Evwiekpaefe A. A comparative study of YOLOv5 and YOLOv7 object detection algorithms. J Comput Soc Inf. 2023;2(1):1-12. doi:10.33736/jcsi.5070
9. Woo S, Park J, Lee JY, Kweon IS. CBAM: Convolutional block attention module. Proc Eur Conf Comput Vis. 2018:3–19. doi:10.1007/978-3-030-01234-2_1
10. Wang L, Zhang X, Fachuan Z, *et al*. Fuzzy-NMS: Improving 3D object detection with fuzzy classification in NMS. arXiv; 2023. https://arxiv.org/abs/2310.13951
11. Mohandoss T, Rangaraj J. Multi-Object Detection using EnhanceTable 1d YOLOv2 and LuNet algorithms in Surveillance Videos. e-Prime Adv Electr Eng Electron Energy. 2024; 8:100535. doi:10.1016/j.prime.2024.100535
12. Xiao X, Feng X. Multi-Object Pedestrian Tracking Using Improved YOLOv8 and OC-SORT. Sensors. 2023;23(20):8439. doi:10.3390/s23208439
13. Shaikh SA, Chopade JJ, Sardey MP. Real-time multi-object detection using enhanced YOLOv5-7S on multi-GPU for high-resolution video. Int J Image Graph. 2023;24(02):2450019. doi:10.1142/S0219467824500190
14. Hazra S, Mandal S, Saha B, Khatua S. UMTSS: a unifocal motion tracking surveillance system for multi-object tracking in videos. Multimed Tools Appl. 2023;82(8):12401–12422. doi:10.1007/s11042-022-13780-5
15. Thakur N, Nagrath P, Jain R, *et al*. Autonomous pedestrian detection for crowd surveillance using deep learning framework. Soft Comput.2023;27 (14): 9383–9399. doi:10.1007/s00500-023-08289-4
16. Che J, He Y, Wu J. Pedestrian multiple-object tracking based on FairMOT and circle loss. Sci Rep. 2023;13(1):4525. doi:10.1038/s41598-023-31806-2
17. Alagarsamy R, Muneeswaran D. Multi-object detection and tracking using reptile search optimization algorithm with deep learning.

Symmetry. 2023;15(6):1194. doi:10.3390/sym15061194
18. Gao X, Wang Z, Wang X, Zhang S, Zhuang S, Wang H. DetTrack: An algorithm for multiple object tracking by improving occlusion object detection. Electronics. 2024;13(1):91. doi:10.3390/electronics13010091
19. Wang J, Hu F, Abbas G, Albekairi M, Rashid N. Enhancing image categorization with the quantized object recognition model in surveillance systems. Expert Syst Appl. 2024;238:122240. doi:10.1016/j.eswa.2023.122240
20. Sahoo PK, Panda MK, Panigrahi U, *et al*. An improved VGG-19 network induced enhanced feature pooling for precise moving object detection in complex video scenes. IEEE Access. 2024;12:45847–64. doi: 10.1109/ACCESS.2024.3381612
21. Arunnehru J. Deep learning-based real-world object detection and improved anomaly detection for surveillance videos. Mater Today Proc. 2023;80:2911–2916. doi:10.1016/j.matpr.2021.07.064
22. Li A, Sun S, Zhang Z, Feng M, Wu C, Li W. A multi-scale traffic object detection algorithm for road scenes based on improved YOLOv5. Electronics. 2023;12(4):878. doi:10.3390/electronics12040878
23. Ganivada A, Yara S. A novel deep convolutional encoder–decoder network: Application to moving object detection in videos. Neural Comput Appl. 2023;35:22027–41. doi:10.1007/s00521-023-08956-5
24. Xu R, Wang Z, Liu Z, *et al*. Histopathological tissue segmentation of lung cancer with bilinear CNN and soft attention. Biomed Res Int. 2022;2022:7966553.
25. Nagaraju M, Babu BS, Sai Somayajulu MV, *et al*. An accurate foreground moving object detection based on segmentation techniques and optimal classifier. Concurr Comput Pract Exp. 2022;34(5):e6689.
26. Grandini M, Bagli E, Visani G. Metrics for multi-class classification: an overview. arXiv; 2020. https://arxiv.org/abs/2008.05756
27. Arkin E, Yadikar N, Xu X, Ubul K. A survey: object detection methods from CNN to transformer. Multimed Tools Appl. 2023;82:21353–21383. doi:10.1007/s11042-022-13801-3
28. Wang M, Leelapatra W. A review of object detection based on convolutional neural networks and deep learning**.** Int Sci J Eng Technol (ISJET)*.* 2022;6(1):1–7.

How to Cite: Ahila MM, Rajesh S, Marimuthu J. Multi Object Detection System for Video Surveillance Using an Improved Yolov5 Model. Int Res J Multidiscip Scope. 2026; 7(1): 1040-1049.
DOI: 10.47857/irjms.2026.v07i01.07685