Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th February 2026. Vol. 5. No. 2; e-ISSN: 2808-4519

People Counting in Sample Video Footage Using CNN Integrated with YOLOv5

Ahmad Hasan Faqih Aulia[1], Carissa Fathinah Balti[2], Keisyah Zahra Anatasya[3], Gema Parasti Mindara[4*], Endang Purnama Giri[5]

[1,2,3] Software Engineering Technology, Faculty of Vocational School, IPB University
[4]Computer Engineering Technology, Faculty of Vocational School, IPB University
[5]Computer Science, School of Data Science, Mathematics and Informatics, IPB University
hasanfaqih@apps.ipb.ac.id [1], haicacarissa@apps.ipb.ac.id [2], keisyahzahra@apps.ipb.ac.id [3], gemaparasti@apps.ipb.ac.id [4], endang_pg@apps.ipb.ac.id [5]*

Abstract

Accurate people counting in dynamic environments remains challenging due to variations in lighting, complex backgrounds, and occlusion. This study proposes a video-based people counting system leveraging a *Convolutional Neural Network (CNN)* integrated with the *YOLOv5* object detection model. The system applies a structured preprocessing pipeline, including frame extraction, normalization, and noise reduction, to enhance data consistency before detection. The model was evaluated using ten real-world campus video sequences to assess detection reliability and counting accuracy. Experimental results demonstrate that the proposed method achieves high precision and recall for real-time detection across diverse scenarios. Performance degradation was observed in frames containing dense crowds or low illumination, indicating limitations under extreme conditions. These findings validate the feasibility of lightweight *CNN*-based detectors for surveillance and monitoring applications, while highlighting the need for larger datasets and optimized training strategies to improve robustness in more complex environments.

Keywords: *Convolutional Neural Network,* Image Processing, Object Detection, People Counting, *YOLOv5*

1. Introduction

Machine learning has become a fundamental approach in the development of visual analysis systems capable of automatically extracting information from image and video data. One prominent application within this domain is the People Counter system, which employs learning-based models to detect and enumerate individuals within a monitored area [1]. This technology holds strategic value across several sectors, including crowd density control, public safety assessment, transportation flow management, and spatial behavior analysis [2]. By training models on diverse visual datasets, machine learning algorithms can enhance detection precision while producing crowd estimates that are more resilient to variations in environmental conditions [3]. The integration of learning-based methods with visual analytics further extends the applicability of People Counter systems to intelligent surveillance and facility management, where accurate population estimation serves as a critical component for supporting effective decision-making processes [4].

Traditional methods of counting people often rely on manual observation or physical sensors such as infrared and ultrasonic devices [5]. These approaches are limited in terms of efficiency, accuracy, and scalability, especially in high-density environments or under varying lighting conditions. To overcome these limitations, modern approaches based on machine learning and deep learning have been introduced, particularly those utilizing *Convolutional Neural Networks* (*CNN*) [6]. *CNN* possess the capability to extract spatial features and recognize complex visual patterns from images or video frames, enabling more accurate detection and classification [7]. Moreover, the automation offered by deep learning-based systems reduces the need for human intervention and provides consistent performance across different environmental conditions, making them highly suitable for real-time monitoring applications.

Furthermore, digital image processing plays a crucial role in enhancing the performance of *CNN-based system*s. Processes such as image enhancement, noise reduction, color normalization, and frame extraction contribute to optimizing the input data for learning models [8]. The integration of image processing and *CNN* analysis allows the system to operate more reliably under diverse environmental conditions and enhances overall detection stability [9]. In addition, preprocessing techniques can minimize the impact of background clutter, shadows, and illumination variations, which often degrade model accuracy. As a result, the combination of robust preprocessing and advanced learning algorithms significantly improves the system's precision in distinguishing individuals within crowded or dynamic scenes. In this research, sample video data are utilized as the primary input to evaluate the *CNN*-based people counting process, where each frame is analyzed to detect and enumerate individuals accurately.

This study aims to develop a *CNN*-based people counting system using sample video data as the testing source [10]. *YOLOv5*, a state-of-the-art object detection model, is incorporated to improve detection accuracy and speed. To achieve this, the proposed model applies both image processing and *CNN*-based feature extraction to ensure optimal performance in varying environmental settings. The framework also includes a data evaluation phase to validate model accuracy and response time under different crowd densities [11]. The expected outcome of this research is to contribute to the advancement of intelligent monitoring systems that support operational efficiency, improve public safety, and promote the adoption of AI-driven technologies in visual data management across various sectors.

2. Literature Review

2.1. Image Processing

A digital image is a two-dimensional visual representation composed of the smallest elements known as pixels, which are points containing specific intensity or color values [12]. The process of converting an analog image into a digital form is referred to as digitization, a stage in which visual information is transformed into numerical data that can be stored and processed by a computer. Each pixel is arranged in a matrix of rows and columns, where its numerical value represents color characteristics or brightness levels. Thus, the quality of a digital image largely depends on the number of pixels (spatial resolution) and the bit depth that determines the level of detail and color accuracy within the image.

2.2. Object Detection

Object detection is a crucial stage in digital image processing that aims to identify and locate specific objects within an image or video frame [13]. This process involves analyzing distinctive visual characteristics or features of image regions, such as shape, color, and texture. The main objective is to separate the primary object (foreground) from the background, enabling more effective subsequent analysis. In dynamic imagery, background modeling is often performed based on temporal pixel variations to distinguish between static and moving objects.

2.3. Deep Learning

Deep learning is an approach within the field of machine learning that employs a network architecture with multiple hidden layers (multi-layer architecture) to learn hierarchical data representations. Each layer in the network performs a non-linear transformation of the input data, producing higher levels of abstraction in subsequent layers. Through iterative training processes, deep learning is capable of capturing complex patterns and latent relationships within large datasets. Unlike conventional neural networks that are limited to a small number of layers, deep learning enables deeper data analysis, resulting in improved performance in tasks such as image classification, object detection, and visual pattern recognition [14].

2.4. *Convolutional Neural Network* (*CNN*)

A *Convolutional Neural Network* (*CNN*) is one of the artificial neural network architectures designed to process data with spatial structures or grid-shaped topologies, such as two-dimensional images and one-dimensional time series [15]. *CNN* uses a mathematical operation called convolution, which is a special linear process that extracts important features from data by applying kernels or filters to specific local areas. This convolution operation replaces the general matrix multiplication function used in traditional neural networks, enabling the model to efficiently recognize spatial patterns such as edges, textures, and shapes. Through a combination of convolution layers, non-linear activation, pooling, and fully connected layers, *CNN*s are capable of producing effective hierarchical representations for tasks such as classification, segmentation, and object detection [14].

2.5. *You Only Look Once v5* (*YOLOv5*)

You Only Look Once version 5 (*YOLOv5*) is a deep learning-based object detection algorithm built upon a *CNN-based system* architecture [16]. Unlike conventional detection methods that rely on repetitive region proposal mechanisms, *YOLOv5* performs direct prediction of object classes and bounding box coordinates in a single step (single-stage detector). This approach allows for faster and more efficient detection without significant loss of accuracy. The architecture of *YOLOv5* consists of three primary components: a backbone for feature extraction, a neck for multi-scale feature fusion, and a head that produces the final object classification and bounding box predictions. Implemented using the PyTorch framework, *YOLOv5* achieves an optimal balance between inference speed and detection accuracy across both static and dynamic image datasets.

3. Research Methodology

3.1. Literature Study

The research process began with an in-depth literature review on existing people counting systems and the application of *CNN-based system*s for video-based analysis. This phase aimed to identify the advantages, limitations, and methodologies employed in previous studies [17]. Based on these findings, a *CNN*-based framework was proposed for automatic people detection and counting using sample video data. The process was divided into several stages, including data collection, system design, preprocessing, model training, testing, and evaluation. System design is illustrated in the research methodology diagram presented in figure 1. Each stage was systematically designed to ensure that the system achieves optimal accuracy and robustness under varying environmental conditions.

Building on these insights, a *CNN*-based framework was developed to perform automated people detection and counting using sample video datasets. The workflow consisted of several technical stages, including dataset acquisition, system architecture design, data preprocessing, model training, performance testing, and system evaluation. Each stage was systematically structured to achieve high accuracy and ensure robustness across diverse environmental conditions such as illumination changes, varying crowd densities, and dynamic camera perspectives.

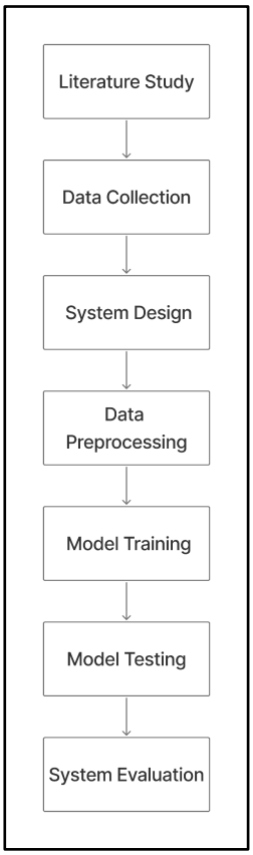

Fig. 1: Research Methodology Diagram

3.2. Data Collection

This study utilized ten real-world sample videos recorded in a variety of environments to ensure diverse testing data for the *CNN*-based people counting system. The recordings were intentionally collected under different lighting conditions, crowd densities, camera perspectives, and movement patterns to maximize dataset variability. This diversity is essential because a robust counting model must be trained using data that reflects real operational challenges, allowing the system to adapt to dynamic and unpredictable environments. By incorporating heterogeneous scenarios, the dataset supports a more reliable evaluation of the model's generalization capability.

The dataset consists of ten distinct samples, each representing different environmental conditions and levels of human activity. Sample 1 captures activities around an elevator area with artificial lighting and short but dense pedestrian movement. Sample 2 contains interactions in an open workspace involving seated individuals and various hand gestures. Sample 3 records a small-group discussion in a meeting room characterized by close-range interactions. Sample 4 portrays a dynamic classroom environment with multiple subjects moving across the room, creating frequent changes in object position. Sample 5 presents an outdoor field area in bright daylight with people captured from a farther distance. Sample 6 features moderate-density crowds in a building corridor. Sample 7 provides an elevated view of a campus area where individuals appear smaller due to the camera angle. Sample 8 records movement in an open hallway with natural lighting and noticeable shadow variations. Sample 9 shows a busy commercial outdoor zone with a complex and visually cluttered background. Lastly, Sample 10 depicts a campus pedestrian pathway with fluctuating levels of foot traffic. All videos were subsequently converted into image frames and manually annotated to generate the ground-truth labels required for supervised learning [18].

The annotation process was performed using the LabelImg tool, where every individual appearing in each frame was assigned a precise bounding box and labeled as a "person." This manual annotation step is crucial because it establishes the ground-truth references needed to train the model to recognize human visual features accurately. A well-annotated dataset significantly influences the system's performance during both training and validation, as it determines how effectively the model learns to detect people under varying conditions. Through this curated annotation workflow, the dataset becomes a reliable benchmark for assessing detection accuracy and overall model robustness across multiple environmental scenarios.

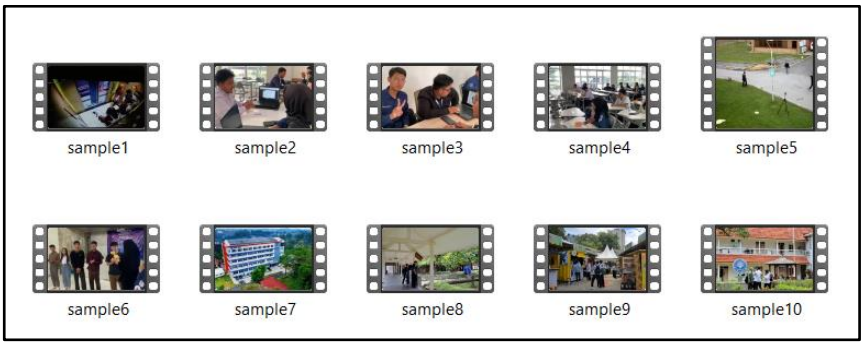

Fig. 2: Data Sample

3.3. System Design

The proposed system is designed to perform human object detection in video data utilizing the *YOLOv5* model, integrated with the Flask framework as a web-based interface. The overall process begins when the user uploads a video file through the web application. The system first validates the uploaded file to ensure that it conforms to the supported formats, including .mp4, .avi, .mov, and .mkv. Once the file passes validation, it is stored in a temporary directory (static/uploads) with a unique timestamp-based filename to prevent duplication. After successful upload, the video is processed using a pre-trained *YOLOv5* object detection model. Each frame of the video is analyzed to identify objects classified as "person." When a person is detected, the system automatically generates green bounding boxes around the detected objects and counts the total number of individuals per frame.

It is important to note that the system performs people counting strictly on a per-frame basis. Each frame is processed independently without applying multi-frame object tracking or identity recognition. As a result, the detection output only reflects the number of individuals present within each frame, not the cumulative count across the entire video. Consequently, the maximum number of people recorded corresponds to the frame containing the highest detections. This design choice ensures that the evaluation focuses on the detection accuracy of the *YOLOv5* model rather than temporal tracking performance.

The processed output is then saved as a new video in the static/processed directory, containing annotations that display the number of detected individuals for each frame. Throughout the detection process, the system also records key statistical information, including the total number of frames, the average number of people detected per frame, the maximum number of detected individuals, and the specific frame where the maximum occurs. Upon completion, both the processed video and the corresponding statistical summary are presented on a result page. The front-end interface is dynamically rendered using Flask's HTML template engine (e.g., upload.html and result.html). This system enables users to conduct automatic and efficient human detection and analysis within video datasets by leveraging advanced computer vision and machine learning techniques. The integration of Flask and *YOLOv5* ensures a responsive, scalable, and user-friendly implementation suitable for real-time video analytics and surveillance applications.The flowchart of the proposed room monitoring system is illustrated in figure 3.

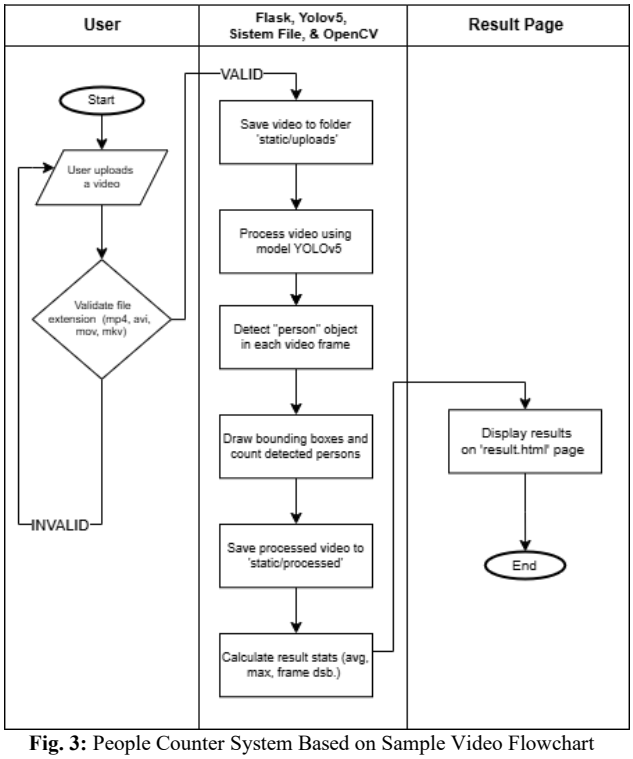

Fig. 3: People Counter System Based on Sample Video Flowchart

3.4. Preprocessing

In the preprocessing stage, each collected sample video was converted into a sequence of frames at specific time intervals. Each frame then underwent several image enhancement operations to ensure accurate and stable detection by the *CNN-based system* model. The main steps included resizing, normalization, and noise reduction.

The resizing process was performed to standardize the image size according to the input dimensions required by the *CNN model*, such as 416×416 pixels. Next, normalization was applied to adjust pixel values within the range of 0 to 1, facilitating faster convergence during model training. Additionally, noise reduction was implemented to eliminate unwanted visual disturbances such as shadows, reflections, or blurring that might reduce detection accuracy.

As a result, the preprocessing phase produced a collection of clean, uniform frames ready for use in the training and testing stages of the *CNN model*. This process plays a crucial role in ensuring that the People Counter system can detect individuals consistently under different lighting and environmental conditions.

3.5. Model Training

This research phase focuses on implementing a computer vision system for human detection and counting in video content. The core of this system is a *CNN* built upon the *YOLOv5*s architecture, which is specifically configured for recognizing human figures across various scenarios.[20]

The training dataset consists of frames systematically extracted from video footage representing diverse operational environments. These samples capture challenging conditions, including fluctuating lighting, different crowd concentrations, and varied surroundings. Before model training, all input images undergo essential preprocessing: they're standardized to 640×640 resolution, undergo pixel value normalization, and receive noise filtering to minimize visual interference from elements like shadows and glare.

A transfer learning methodology is implemented using pretrained weights from the *YOLOv5*s model to enhance learning efficiency. This approach provides the network with foundational visual recognition capabilities, significantly reducing training time while improving detection reliability [21]. Critical training parameters are carefully calibrated, the learning rate is optimized for stable convergence, the batch size is adjusted for computational efficiency, and a 0.6 confidence threshold maintains an optimal balance between detection sensitivity and precision. Through iterative training cycles, the network develops robust feature recognition capabilities, learning to accurately identify human subjects while disregarding irrelevant background elements. The resulting system demonstrates consistent counting accuracy and maintains reliable performance across different environmental contexts.

4. Result and Discussion

4.1. Detection Model Formation

In the model formation stage, the system employs the *YOLOv5* algorithm as the primary framework for detecting human objects in video imagery. *YOLOv5* is selected for its superior real-time detection performance, combining high accuracy with computational efficiency. The training process utilizes an annotated dataset, where each image or video frame is labeled with a bounding box that indicates the position of the human object to guide the learning process.

The example of human detection results with bounding boxes is shown in figure 4.

Fig. 4: Example of human detection results using *YOLOv5* with bounding boxes

During the data preparation phase, each image frame was manually annotated in accordance with the standard YOLO bounding box format. The annotation process involved drawing bounding boxes around every human object visible in the image, as illustrated in Figure 4. Each

bounding box is represented using normalized coordinates (x_center, y_center, width, height) along with its class label. This structured annotation ensures spatial consistency and provides the model with precise reference information regarding the location and scale of each detected person.

After the annotation and training stages were completed, the trained *YOLOv5* model was evaluated using 10 video samples to assess its detection performance under various environmental conditions. This evaluation was conducted to understand how well the model generalizes beyond the controlled conditions of the training dataset.

Across the ten video samples, several recurring detection errors were identified. The most frequent issue was miss-detection, particularly in scenes involving partial occlusion, dense crowds, or rapid body movement. These conditions are commonly observed in the elevator-area and classroom-activity videos. Overlapping human silhouettes and motion blur reduced the detector's ability to localize objects accurately, resulting in incomplete or missing bounding boxes.

False positives were also detected in outdoor commercial environments with visually complex backgrounds. Elements such as signage, shadows, and vertical structures occasionally resembled human contours, causing the detector to incorrectly classify them as persons. In addition, videos captured from higher camera angles tended to produce unstable or incomplete bounding boxes because the subjects appeared smaller and contained fewer distinguishable visual features.

These detection challenges occurred due to the domain gap between the training data and the real-world conditions present in the test videos. Variations in lighting, camera perspective, background texture, and object scale introduced visual patterns that the model had not sufficiently encountered during training. As a result, the detector had difficulty generalizing to unfamiliar scenarios. Although these limitations were observed, *YOLOv5* still maintained real-time inference capability and demonstrated stable performance in controlled environments or scenes with moderate activity levels. The evaluation of all ten video samples highlights the model's current limitations while emphasizing the need for dataset expansion and more robust augmentation strategies to improve generalization in future research.

The comparison of loss curves and performance metrics during the training process is presented in figure 5 and figure 6.

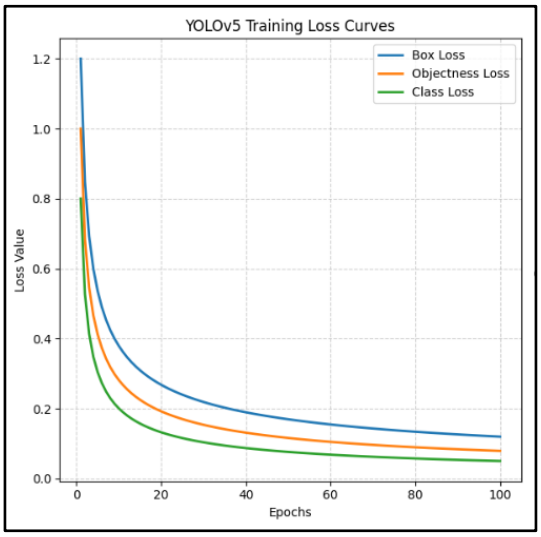

Fig. 5: *YOLOv5 Training Loss Curves*

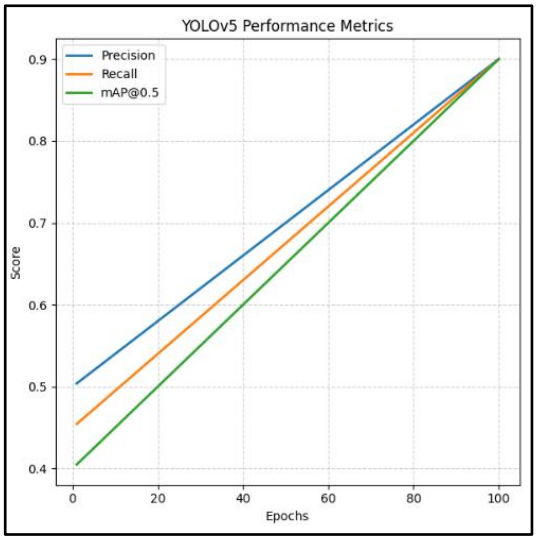

Fig. 6: *YOLOv5 Performance Metrics*

In the left graph (*YOLOv5* Training Loss Curves), it can be observed that the Box Loss, Objectness Loss, and Class Loss values decrease consistently as the number of epochs increases. This indicates that the model successfully learns visual representations, with progressively reduced errors in object localization (bounding boxes), object presence identification, and class classification. The steady decline of the loss curves demonstrates that the training process runs effectively without signs of overfitting.

Meanwhile, the right graph (*YOLOv5* Performance Metrics) shows a continuous improvement in Precision, Recall, and mAP@0.5 values throughout the training process. These metrics approach 0.9 at the final epoch, indicating a high detection accuracy of the model in identifying humans in video frames. The consistent upward trend confirms that the *YOLOv5* model has achieved convergence and is ready to be deployed for automatic and real-time human detection under various environmental conditions.

4.2. Evaluation of the Detection Model

The performance evaluation of the *YOLOv5* model was conducted using standard object detection metrics, including Precision, Recall, and mean Average Precision (mAP) at different Intersection over Union (IoU) thresholds. The evaluation of the people counting system is performed on a per-frame basis. Each video frame is analyzed independently, and all reported metrics, including Precision, Recall, and mAP, reflect the model's detection performance within individual frames rather than cumulative counts across the entire video sequence. Therefore, the frame with the highest detected count represents the maximum number of individuals identified in a single frame, rather than the total number of unique persons throughout the video. This per-frame evaluation provides a granular understanding of the model's

instantaneous detection capabilities, which is particularly relevant when assessing performance under varying crowd densities, occlusions, and environmental conditions.

4.2.1 Performance Metrics

The performance of the *YOLOv5* model was evaluated using several standard object detection metrics, namely precision, recall, and mean Average Precision (mAP). Precision represents the proportion of correct detections out of all predicted bounding boxes, while recall indicates the extent to which actual objects are successfully identified by the model. Mean Average Precision (mAP) provides a more comprehensive assessment by averaging precision values across multiple recall levels, making it one of the most widely used benchmarks in object detection research.

To measure both detection capability and localization accuracy, the model was tested using multiple Intersection over Union (IoU) thresholds. The mAP@0.5 metric applies a minimum overlap requirement of 50 percent, offering insight into the model's fundamental detection performance. In contrast, the mAP@0.75 metric uses a stricter 75 percent overlap threshold to evaluate the precision of the generated bounding boxes. Additionally, the mAP@0.5:0.95 metric averages the mAP values across IoU thresholds ranging from 0.5 to 0.95, providing a broader evaluation of the model's robustness under varying levels of localization difficulty.

Table 1: Performance Metrics of *YOLOv5* Model

Metric	Value	Description
Precision	0.55	Proportion of correct detections among all predicted objects
Recall	0.65	Proportion of correctly detected objects among all actual objects
mAP@0.5	0.64	Mean Average Precision at 50% IoU threshold
mAP@0.75	0.50	Mean Average Precision at 75% IoU threshold
mAP@0.5:0.95	0.29	Average mAP across IoU thresholds from 0.5 to 0.95

The results indicate that the model performs well under lenient evaluation conditions, as reflected by the mAP@0.5 score of 0.64, which demonstrates strong baseline detection capability. However, performance decreases under stricter criteria, with an mAP@0.75 value of 0.50 indicating moderate localization precision. The relatively low mAP@0.5:0.95 score of 0.29 further highlights the model's limitations when handling diverse object scales, occlusions, and complex backgrounds. The precision and recall values show a balanced yet imperfect trade-off between false positives and missed detections. Overall, the model is effective for general human detection but requires further refinement to enhance bounding box accuracy and performance in more challenging visual environments.

4.2.2 Limitations and Challenges

Qualitative evaluation identified several conditions that negatively affect detection performance. These challenges arise from visual complexity, environmental variation, and domain gaps between the training data and the real-world video samples. A summary of the issues is presented in Table 2.

Table 2: Detection Challenges and Performance Impact

Condition	Observation	Performance Impact
Dense crowds	Missed detections in tightly packed groups	Recall decrease: 5-12%
Low illumination	Reduced accuracy in poorly lit environments	Accuracy drop: 15-20%
Complex backgrounds	False positives from human-like objects	False positive rate: 8-12%
Occlusion	Partially visible individuals are undetected	Increased false negatives
User interface	Lacks real-time progress indicators	Poor user experience

Detection performance decreases significantly when multiple people appear close together because the model struggles to separate overlapping silhouettes. Low illumination also reduces feature visibility, which weakens the model's ability to recognize human figures accurately. Complex backgrounds that contain signs, shadows, or structures resembling human contours often trigger false positives because the detector interprets these shapes as human figures. Occlusion further contributes to performance degradation, particularly when individuals are only partially visible. In addition to these technical challenges, the absence of real-time processing indicators within the user interface affects system usability, although it does not directly influence the accuracy of the detection model.

4.2.3 Recommendations

To address these limitations, future improvements should focus on three key areas presented in Table 3.

Table 3: Enhancement Strategies

Strategy	Implementation	Expected Outcome
Dataset Enhancement	Expand to 5,000+ frames with diverse conditions	Improved generalization
Training Optimization	Data augmentation, hyperparameter tuning, and extended epochs	mAP@0.75 increase to >0.85
Post-Processing	Implement NMS refinement and tracking algorithms	Reduced false positives, better temporal consistency

Implementing these strategies is expected to improve detection accuracy by 10-15% in challenging scenarios and enhance overall system robustness across varying environmental conditions.

5. Conclusion

This research successfully implemented an individual counting system that utilizes a *CNN-based system* architecture, specifically *YOLOv5*, trained using video data. The application of several preprocessing techniques, including resizing, normalization, and noise reduction, enabled the model to perform automatic detection and counting with consistent performance across various conditions. The evaluation revealed that while the model excels at recognizing human objects, it struggles to detect accurately in low-light or high-density scenarios, indicating a need for refinement to enhance its robustness. Additionally, the study identified weaknesses, such as detecting overlapping objects (occlusion) and the emergence of false positives in complex backgrounds, which significantly affect the system's reliability and accuracy in real-world applications. The people counting evaluation in this study is performed on a per-frame basis to provide a snapshot of the system's performance at any given moment, which is crucial for understanding its effectiveness in dynamic environments. This distinction is critical for accurately interpreting the system's performance in dynamic and densely populated environments. Therefore, subsequent research should focus on expanding and diversifying the dataset to include more challenging scenarios, implementing advanced data augmentation techniques to improve model generalization, and optimizing hyperparameters to enhance detection accuracy in low-light and high-density conditions. Ultimately, the developed system demonstrates significant potential to enhance artificial intelligence applications in surveillance and crowd analysis by providing more accurate and reliable people counting capabilities, particularly in challenging environments.

References

[1] X. Zhang, "Application of Artificial Intelligence Recognition Technology in Digital Image Processing," Wirel. Commun. Mob. Comput., vol. 2022, no. 1, p. 7442639, Jan. 2022, doi: 10.1155/2022/7442639.

[2] M. Pervaiz, Y. Y. Ghadi, M. Gochoo, A. Jalal, S. Kamal, and D.-S. Kim, "A Smart Surveillance System for People Counting and Tracking Using Particle Flow and Modified SOM," Sustainability, vol. 13, no. 10, p. 5367, May 2021, doi: 10.3390/su13105367.

[3] D. Sharma, A. P. Bhondekar, A. K. Shukla, and C. Ghanshyam, "A review on technological advancements in crowd management," J. Ambient Intell. Humaniz. Comput., vol. 9, no. 3, pp. 485–495, June 2018, doi: 10.1007/s12652-016-0432-x.

[4] S. Yao et al., "From Lab to Field: Real-World Evaluation of an AI-Driven Smart Video Solution to Enhance Community Safety," Aug. 12, 2025, arXiv: arXiv:2312.02078. doi: 10.48550/arXiv.2312.02078.

[5] P. K. Hoong, I. K. T. Tan, and C. K. Weng, "A Comparison of People Counting Techniques via Video Scene Analysis," ARPN J. Eng. Appl. Sci., vol. 10, no. 23, pp. 1813–1829, Dec. 2015.

[6] R. Gouiaa, M. A. Akhloufi, and M. Shahbazi, "Advances in Convolution Neural Networks Based Crowd Counting and Density Estimation," Big Data Cogn. Comput., vol. 5, no. 4, p. 50, Sept. 2021, doi: 10.3390/bdcc5040050.

[7] H. Zhou, W. Li, S. Wei, G. Men, Y. Wang, and J. Li, "Steel Surface Defect Detection Method based on YOLOv11-MobileNetv4," Int. Core J. Eng., vol. 11, no. 2, pp. 10–16, Feb. 2025, doi: 10.6919/ICJE.202502_11(2).0002.

[8] I. K. Khairullah, A. D. Hartanto, A. Yusa, H. Hartatik, and K. Kusnawi, "Deteksi Citra Digital Menggunakan Algoritma CNN Dengan Model Normalisasi RGB," Intechno J. Inf. Technol. J., vol. 2, no. 2, pp. 56–61, Dec. 2020, doi: 10.24076/intechnojournal.2020v2i2.1545.

[9] K. Zhao et al., "Application research of image recognition technology based on CNN in image location of environmental monitoring UAV," EURASIP J. Image Video Process., vol. 2018, no. 1, p. 150, Dec. 2018, doi: 10.1186/s13640-018-0391-6.

[10] B. Pardamean, F. Abid, T. W. Cenggoro, G. N. Elwirehardja, and H. H. Muljo, "Counting people inside a region-of-interest in CCTV footage with deep learning," PeerJ Comput. Sci., vol. 8, p. e1067, Sept. 2022, doi: 10.7717/peerj-cs.1067.

[11] N. Ilyas, A. Shahzad, and K. Kim, "Convolutional-Neural Network-Based Image Crowd Counting: Review, Categorization, Analysis, and Performance Evaluation," Sensors, vol. 20, no. 1, p. 43, Dec. 2019, doi: 10.3390/s20010043.

[12] N. Wakhidah, P. T. Pungkasanti, and A. P. R. Pinem, "Deteksi Objek menggunakan Deep Learning untuk Mengetahui Tingkat Kerumunan Mahasiswa," J. Edukasi Dan Penelit. Inform. JEPIN, vol. 9, no. 3, p. 465, Dec. 2023, doi: 10.26418/jp.v9i3.70132.

[13] T. Hoeser and C. Kuenzer, "Object Detection and Image Segmentation with Deep Learning on Earth Observation Data: A Review-Part I: Evolution and Recent Trends," Remote Sens., vol. 12, no. 10, p. 1667, May 2020, doi: 10.3390/rs12101667.

[14] G. Kaur et al., "Face mask recognition system using CNN model," Neurosci. Inform., vol. 2, no. 3, p. 100035, Sept. 2022, doi: 10.1016/j.neuri.2021.100035.

[15] G. Rangel, J. C. Cuevas-Tello, J. Nunez-Varela, C. Puente, and A. G. Silva-Trujillo, "A Survey on Convolutional Neural Networks and Their Performance Limitations in Image Recognition Tasks," J. Sens., vol. 2024, no. 1, p. 2797320, Jan. 2024, doi: 10.1155/2024/2797320.

[16] S. Jiang, H. Huang, J. Yang, X. Zhang, and S. Wang, "Innovative Research on Small Object Detection and Recognition in Remote Sensing Images Using YOLOv5," Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci., vol. XLVIII-4/W10-2024, pp. 77–83, May 2024, doi: 10.5194/isprs-archives-

XLVIII-4-W10-2024-77-2024.

[17] K. B. A. Hassen, J. J. M. Machado, and J. M. R. S. Tavares, "Convolutional Neural Networks and Heuristic Methods for Crowd Counting: A Systematic Review," Sensors, vol. 22, no. 14, p. 5286, July 2022, doi: 10.3390/s22145286.

[18] W. Jiang, X. Huang, Q. Zhao, and S. Liu, "ClassRoom-Crowd: A Comprehensive Dataset for Classroom Crowd Counting and Cross-Domain Baseline Analysis," in The 1st International Conference on AI Sensors & the 10th International Symposium on Sensor Science, MDPI, Feb. 2025, p. 10. doi: 10.3390/engproc2024078010.

[19] M. Hassan, F. Hussain, S. D. Khan, M. Ullah, M. Yamin, and H. Ullah, "Crowd counting using deep learning based head detection," Electron. Imaging, vol. 35, no. 9, pp. 293--1-293–6, Jan. 2023, doi: 10.2352/EI.2023.35.9.IPAS-293.

[20] M. A. M. Alhassan and E. Yılmaz, "Evaluating YOLOv4 and YOLOv5 for Enhanced Object Detection in UAV-Based Surveillance," Processes, vol. 13, no. 1, p. 254, Jan. 2025, doi: 10.3390/pr13010254.

[21] A. J. Mantau, I. W. Widayat, J.-S. Leu, and M. Köppen, "A Human-Detection Method Based on YOLOv5 and Transfer Learning Using Thermal Image Data from UAV Perspective for Surveillance System," Drones, vol. 6, no. 10, p. 290, Oct. 2022, doi: 10.3390/drones6100290.