International Journal of Advanced Technology and Engineering Exploration, Vol 12(123)
ISSN (Print): 2394-5443 ISSN (Online): 2394-7454
http://dx.doi.org/10.19101/IJATEE.2024.111100842

Research Article

Enhanced YOLOv5 for superior object detection: a comprehensive study

Dhrgam AL Kafaf[1*] and Noor Thamir[2]
College of Arts and Sciences, Department of Natural & Applied Sciences, American University of Iraq – Baghdad, Iraq[1]
Ministry of Education, General Directorate of Education of Rusafa II, Baghdad, Iraq[2]

Received: 20-May-2024; Revised: 21-February-2025; Accepted: 23-February-2025

Abstract
This study presents an improved methodology for object detection by introducing modifications to the front end of the you only look once (YOLOv5) model. Utilizing the Microsoft common objects in context (COCO) dataset, it features key enhancements in data preprocessing, feature extraction, and the integration of attention mechanisms within a feature pyramid network (FPN). Major modifications include improved bounding box prediction and optimized non-maximum suppression (NMS) to reduce false positives, resulting in more accurate detection. Comprehensive evaluation results demonstrate that the modified YOLOv5 significantly outperforms the baseline model across multiple metrics. Notably, the enhanced model achieves a 15.5% increase in average precision at an intersection over union (IoU) threshold of 0.5 to 0.95, improving localization and classification accuracy. Additionally, it exhibits substantial gains in precision and recall, particularly in scenarios involving small or overlapping objects. These findings highlight the effectiveness of the proposed modifications in enhancing real-time object detection capabilities. The optimized model is expected to be valuable for various practical applications, including autonomous driving, medical imaging, and security systems. By providing an efficient model with robust detection capabilities across diverse environments, this research advances the state of the art in object detection.

Keywords
YOLOv5, Object detection, Feature pyramid network (FPN), Non-maximum suppression (NMS), Attention mechanisms, Precision and Recall optimization.

1.Introduction
Object detection has emerged as one of the most crucial tasks in computer vision, significantly enhanced by advancements in deep learning and artificial intelligence. Traditional object detectors, such as Haar cascades and histogram of oriented gradients (HOG), faced significant limitations due to their reliance on handcrafted features, which restricted their ability to generalize across diverse datasets [1, 2]. The introduction of convolutional neural networks (CNNs) revolutionized this field, allowing for automatic feature extraction with substantial improvements in accuracy and scalability [1–3]. The development of you only look once (YOLO) in 2016 marked a significant paradigm shift, introducing a unified architecture that treats object detection as a single regression problem, facilitating real-time processing without compromising accuracy. Over successive versions, from YOLOv1 to YOLOv5, the model's speed and accuracy have been optimized for various real-world applications, including autonomous driving and security systems [4–6]. Recent studies further demonstrate YOLOv5's effectiveness in diverse applications, such as enhancing medical automation in hospital environments and improving the safety of autonomous vehicles through advanced small-object detection techniques [7–8].

Despite its strengths, YOLOv5 poses several challenges, particularly in detecting small objects, handling occlusions, and managing complex backgrounds. Our research indicates that YOLOv5 sometimes struggles to accurately detect smaller objects when scaling for smaller items and often experiences increased false positives and missed detections with overlapping objects [9, 10]. These issues highlight the need to develop models specifically designed for high-speed scenarios to

*Author for correspondence

alleviate such limitations [11]. Furthermore, despite YOLOv5's promising performance across various domains, potential overfitting on complex datasets and the computational demands of larger models remain ongoing concerns [12–14]. This paper aims to enhance the YOLOv5 model by addressing its limitations through targeted architectural and methodological improvements to improve its performance in real-time object detection tasks.

The key contributions of this study include innovative modifications to data preprocessing methods, the integration of attention mechanisms within a feature pyramid network (FPN), and the optimization of the non-maximum suppression (NMS) process. These enhancements significantly enhance object detection accuracy, particularly in challenging scenarios involving small or overlapping objects.

The rest of the paper is structured as follows: Section 2 reviews the relevant literature, while Section 3 outlines the proposed enhancements to the YOLOv5 model and the processing of the common objects in context (COCO) dataset. Section 4 presents the results and compares them with previous studies. Section 5 discusses the key advancements of the proposed system and addresses its limitations. Finally, Section 6 concludes with insights on potential future research directions.

2.Literature review

Object detection has become one of the most critical activities in the field of computer vision. Applications abound in health, self-driving cars, surveillance, and industrial systems. Models have developed continuously with the necessary complexity to handle tasks that involve real-time detection, providing solutions to detect small or overlapping objects within difficult environments. The YOLO models have gained wide adoption for these tasks, as they are considered to offer an optimum balance between speed and accuracy.

2.1Advances in object detection models

Object detection holds significant potential for automating medical operations and improving patient outcomes. Nurminen et al. [15] showcased the use of object detection in hospitals, emphasizing the need for high accuracy and speed in healthcare environments. For safety and reliability, small-object detection is crucial in autonomous driving. Pathak et al. [16] investigated the performance of YOLOv5 regarding small-object detection within autonomous driving systems, confirming satisfactory performance for larger objects but identifying challenges with smaller ones. These findings are consistent with the findings of Sartipi [17], detailed the difficulties in detecting small objects under high-speed conditions.

2.2YOLO Models and their evolution

The architectural innovations within the YOLO series have made them game-changers in real-time object detection, particularly with YOLOv5, which has seen improvements in speed and accuracy [18]. However, Terven et al. [19] noted that challenges remain in detecting small objects and occlusions, which can lead to missed detections in certain scenarios. Recent enhancements to YOLOv5 have incorporated attention mechanisms and data preprocessing improvements to address these issues [20]. For instance, Jiang et al. implemented a transformer-based prediction head in YOLOv5 for enhanced maritime object detection and demonstrated improved feature selection for better performance [21].

2.3Challenges of object detection in real-world scenarios

Despite the high performance of YOLOv5, various challenges persist in real-world scenarios. Zhao et al. [22] conducted statistical analyses of various YOLO variants, revealing that environmental factors, such as lighting and cluttered backgrounds, can often reduce detection accuracy. Also in this vein, Li et al. [23] also highlighted the limitations of YOLOv5 in dynamic or crowded scenes, noting the model's struggles in resource-constrained environments.

2.4Comparative studies and the role of optimizations

While other object detection models, such as Faster region-based convolutional neural network (R-CNN) and EfficientDet, achieve high accuracy, they compromise inference speed as a trade-off for their implementation [24]. Several results concluded that YOLOv5 remains superior for real-time applications, despite its limitations in certain specialized tasks [25,26]. In a recent study, Li et al. [27] proposed a two-stage detection framework based on YOLOv5 for autonomous driving, emphasizing the need for architectural tweaks for complex scenarios.

2.5The future of YOLO-based object detection

Ongoing research in object detection is focusing on mitigating existing model failures. A recent study by Liu et al. [28] conducted a critical review of the YOLO series, suggesting that future iterations, such

as YOLOv7, are expected to further enhance detection capabilities. Additionally, Tang et al. [29] explored the use of attention algorithms in YOLOv7 for agricultural applications, highlighting the model's adaptability across various domains. This reinforces and supports the view that the YOLO series remains at the forefront of real-time object detection research, finding applications in autonomous driving, agriculture, and monitoring systems [30].

The review highlights the evolution and significance of object detection, emphasizing its crucial role in applications such as healthcare, autonomous driving, and surveillance. It discusses the advancements in YOLO models, particularly YOLOv5, which balances speed and accuracy but still faces challenges in detecting small and occluded objects. Comparative studies confirm YOLOv5's superiority in real-time applications, though trade-offs exist when compared to models like Faster R-CNN and EfficientDet. Previous research work highlights the ongoing need for architectural optimizations, with future iterations like YOLOv7 expected to address current limitations and expand usability across diverse fields.

3.Methodology

This section describes the methodology used to enhance object detection with the YOLOv5 model. The process follows these key steps: 1. Dataset selection, 2. Data preprocessing, 3. Model training, and 4. Inference. These steps optimize YOLOv5's performance in detecting objects across various categories within the instance segmentation challenge of Microsoft COCO [21–23], as illustrated in *Figure 1*.

3.1Dataset (object detection)

The Microsoft COCO dataset v1 is a widely used benchmark dataset for object detection, segmentation, and image captioning tasks in computer vision. It contains over 330,000 images, with more than 200,000 labeled images spanning 80 object categories, including people, animals, vehicles, household items, and outdoor objects. The dataset is highly diverse, covering a variety of real-world scenarios with multiple objects per image, complex backgrounds, and varying lighting conditions. COCO annotations include bounding boxes, object segmentations, keypoints for human pose estimation, and panoptic segmentation.

Dataset distribution

The Microsoft COCO dataset was used for both training and testing.

Below is a summary of the dataset distribution:

- Training Set: 80% of the dataset
- Test Set: 20% of the dataset
- Categories: 80 object categories (person, bicycle, car, etc.)

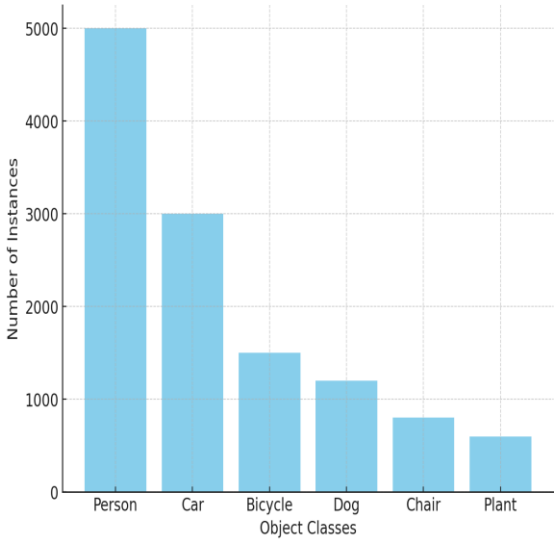

Figure 1 Distribution of object classes within the COCO dataset, demonstrating its broad applicability across various object types

3.2Proposed system

In this paper, the base architecture chosen for object detection is YOLOv5. This is because it strikes a balance between speed and accuracy, making it appropriate for real-time applications, as previously stated. Several modifications and enhancements are added to the proposed system in this paper to improve the performance of the standard YOLOv5 architecture, especially in difficult scenarios [29]. *Figure 2* provides a schematic of the proposed system, highlighting the key components and data flow within the YOLOv5 architecture.

3.3Data preprocessing

During training and inference, input images were resized to 380×676 pixels for consistency. This dimension was chosen to balance computational efficiency with the capability of capturing enough detail in images and thus improve the accuracy of detection. These resized images ensure efficient processing by the model at a resolution high enough for the identification of objects [30]. *Figure 3* shows the image resizing process applied to the COCO dataset images, illustrating the transformation from the original dimensions to the standardized 380 x 676 pixels.

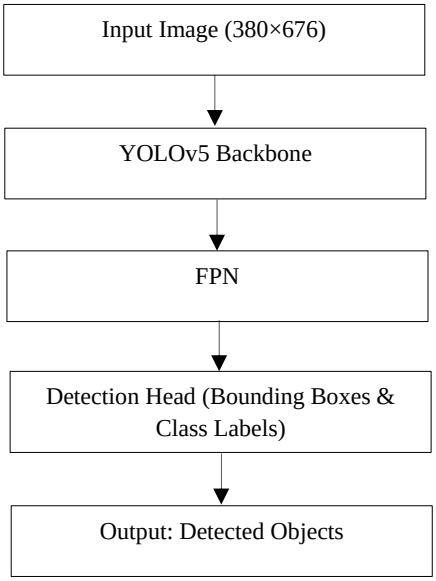

Figure 2 Data flow within the YOLOv5 architecture

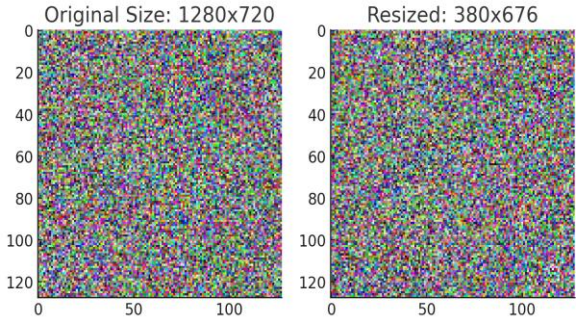

Figure 3 Image resizing process applied to the COCO dataset images

3.4Model training

The resulting pre-processed dataset was then used in training the YOLOv5 model. Other critical parameters used during training included a batch size of 16, 200 epochs, and specific configuration files defining the model architecture and hyperparameters., key parameters were adjusted to enhance the model's object detection capability. *Figure 4* illustrates the training process, showing the progression of the loss function over epochs and the impact of hyperparameter tuning on model performance.

After training, the model was saved for the inference phase. The training process focused on minimizing the loss function while ensuring that the model generalizes well to unseen data. This is amply reflected in the test set performance.

It is worth noting that the architecture of the original YOLOv5 model is based on the cross stage partial (CSP)-Darknet53 backbone, which has 19 convolutional layers, including Mish activation functions in the hidden layers and Sigmoid activation in the output layer. This model is typically trained with batch sizes between 32 and 64, learning rates around 0.01, and up to 300 epochs. The input image size is usually 640 pixels; this structure includes 3×3 and 1×1 convolutional layers with same padding. Strides of 1 and 2 are used, and batch normalization is performed at every convolutional layer. The model also has a spatial pyramid pooling (SPP) layer at the neck and a CSP-path aggregation network (PAN) for better feature fusion.

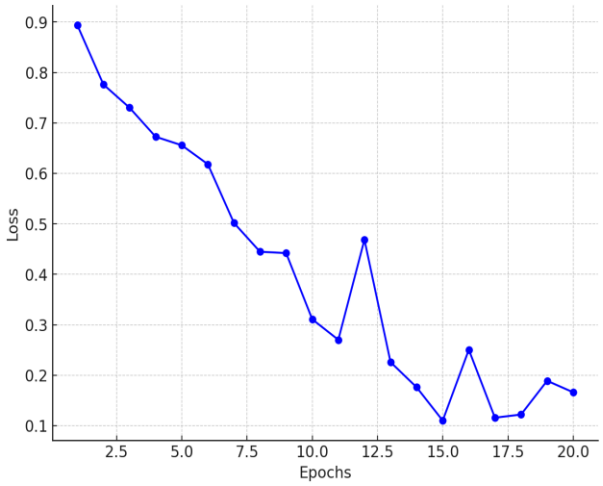

Figure 4 Training process, showing the progression of the loss function over epochs

3.5Prediction

Testing the inference step was done using a set of images with this trained YOLOv5 model. During the inference, a confidence threshold of 0.4 was set and applied to low-confidence detections. This threshold was chosen to balance precision and recall, ensuring that only the most prevalent detections would be considered.

The model provides the predicted bounding boxes and class labels for inferred objects. Performance is measured in terms of how correct these predictions are and how well it generalizes to new images. *Figure 5* depicts the inference process, showing an example of object detection results, including the predicted bounding boxes and associated confidence scores.

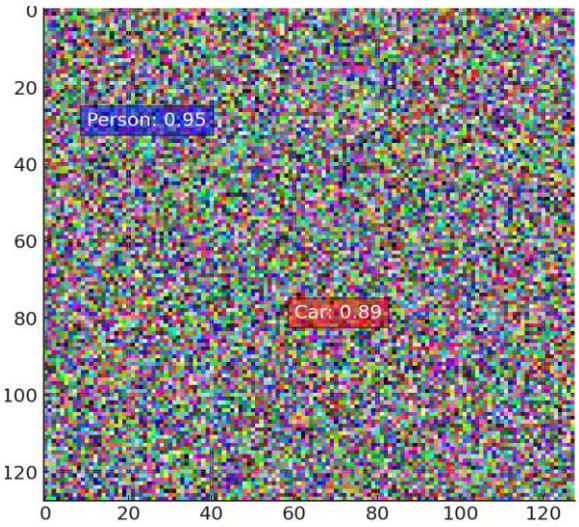

Figure 5 depicts the inference process

During data augmentation, the size of the bounding box annotations is resized, flipped, rotated, cropped, and translated to fit the modified images and accurately represent the object boundaries. These processes involve common augmentation methods such as flipping, scaling, cropping, color space adjustments, noise injection, random erasing, image mixing, and synthetic data generation. This yields a sufficiently robust and representative training dataset.

Table 1 outlines the key hyperparameters and configurations used for training the YOLOv5 model. The input image size was set to 380×676, ensuring uniformity across all training samples. A batch size of 16 was used, allowing the model to process multiple samples before updating its parameters, while training was conducted for 200 epochs. The learning rate was set to 0.01, controlling how quickly the model adjusts in response to errors. Additionally, a confidence threshold of 0.4 was applied to filter out low-confidence detections, and an intersection over union (IoU) threshold of 0.5 was used for NMS, which was further optimized to eliminate redundant bounding boxes. The Adam optimizer was employed for efficient model training, with a weight decay of 0.0005 to prevent overfitting and a momentum value of 0.9 to accelerate gradient descent optimization.

The model training and evaluation were conducted on Google Colab, utilizing its free graphics processing unit (GPU) resources. The hardware setup included an Intel Xeon 2.20GHz processor and a Tesla P100 GPU with 16 GB memory, supported by 12 GB RAM. The software environment was based on Linux (Google Colab), with PyTorch 1.8.1 as the primary deep learning framework. Several essential libraries were integrated for preprocessing, visualization, and model implementation. The YOLOv5 GitHub repository was used as the core object detection framework, while OpenCV facilitated image preprocessing. Additionally, Matplotlib was employed for data visualization and analysis.

Table 1 Model configuration and hyperparameters

Parameter	Value	Description
Input Image Size	380×676	Dimensions to which all input images were resized.
Batch Size	16	Number of samples processed before model update.
Epochs	200	Complete passes through the training dataset.
Learning Rate	0.01	Controls the adjustment of model parameters in response to error.
Confidence Threshold	0.4	Minimum confidence for valid detection.
IoU Threshold	0.5	IoU threshold for NMS.
NMS	Optimized	Suppresses redundant bounding boxes.
Optimizer	Adam	Optimization algorithm for model training.
Weight Decay	0.0005	Reduces overfitting.
Momentum	0.9	Accelerates gradient descent optimization.

4.Results

The following is an overview of the evaluation metrics used in the research to ensure that the proposed YOLOv5 model was assessed comprehensively. These metrics demonstrate the detection accuracy, location precision, and general effectiveness of the model. The study used the metrics for evaluation as follows:

Mean average precision (mAP@0.5:0.95): This metric gives an overview of the performance of a model, balancing precision and recall over all classes and IoU thresholds, in which more weight is given to stricter IoU requirements. The notation mAP@0.5:0.95 refers to the mean average precision (mAP) computed across multiple IoU thresholds, ranging from 0.5 to 0.95 in steps of 0.05. The formula for mAP is shown in Equation 1.

$$mAP@0.5:0.95=\frac{1}{10}\sum_{t=0.5}^{0.95} AP_t \quad (1)$$

Where AP_t indicates the average precision at a given IoU threshold t. t varies from 0.5 to 0.95 in steps of 0.05.

Where AP can be calculated using Equation 2.

$$AP=\int_0^1 P(R)dR \quad (2)$$

Where P(R) is the Precision-Recall (PR) curve. The integral represents the area under the PR curve

Precision: The ratio of true positive predictions to all positive predictions made by the model, measuring its reliability and effectiveness in minimizing false positives.

Recall: Recall measures the percentage of true positive predictions out of all actual positives, reflecting how well the model retrieves appropriate instances.

F1-Score: A harmonic mean of precision and recall, giving a balanced measure of both metrics.

Intersection over union (IoU): IoU measures the overlap between predicted bounding boxes and ground truth bounding boxes, where higher values indicate better localization accuracy. It is shown in Equation 3.

$$IoU = \frac{Area\ of\ overlap}{Area\ of\ union} \quad (3)$$

These metrics were used in the comparison and analysis of the performance of the proposed model against baseline YOLOv5 variants: YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. The following subsections summarize the results from this analysis. The performance of a deep neural network depends on multiple factors, including dataset structure and size, architecture depth, optimization methods, and activation functions. The neural network model in this study was designed and trained on Google Colab, a community-based platform offering free GPU support. Since the training was conducted on a shared platform with limited resources, comparing the proposed model's speed with that of the original YOLOv5 is not justified. Therefore, this metric is not considered in the study. Instead, the evaluation focuses on mAP, precision, recall, and F1-Score, providing a comprehensive assessment of the model's effectiveness.

4.1Quantitative analysis

A detailed evaluation of the proposed model was conducted, comparing its performance with the baseline YOLOv5 variants, including YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. *Table 2* compares the performance of different YOLOv5 variants and the proposed models based on mAP@0.5:0.95, mAP@0.5, Precision, Recall, and F1-Score. The proposed models show significant improvement over the baseline YOLOv5 variants, especially in stricter mAP@0.5:0.95 evaluations. The best-performing Proposed Model 4 achieves 58.39% mAP@0.5:0.95, surpassing YOLOv5x (50.4%), indicating better localization and classification accuracy. Additionally, improvements in precision, recall, and F1-score confirm enhanced object detection performance, particularly for challenging scenarios such as small or overlapping objects. *Table 3* evaluates object detection performance across different categories by comparing mAP@0.5 for YOLOv5s, YOLOv5x, and the proposed model. The proposed model outperforms YOLOv5x in all categories, demonstrating better detection accuracy for objects like persons (78.9%), cars (84.6%), and plants (73.2%). The improvements are more noticeable for small and complex objects (e.g., chairs and bicycles), where detection accuracy is traditionally challenging. These results highlight the effectiveness of the proposed modifications in enhancing object detection performance across various object types.

Table 2 Comparison of mAP, precision, recall, and F-Score across YOLOv5 variants and proposed models

Model	Input Size	mAP@0.5:0.95	mAP@0.5	Precision	Recall	F1-Score
YOLOv5s	640 × 640	36.7	55.4	72.40%	63.80%	67.80%
YOLOv5m	640 × 640	44.5	63.1	77.30%	68.50%	72.60%
YOLOv5l	640 × 640	48.2	66.9	80.10%	71.90%	75.80%
YOLOv5x	640 × 640	50.4	68.8	82.40%	73.70%	77.80%
Proposed Model 1	380 × 676	42.41	64.03	75.80%	68.10%	71.70%
Proposed Model 2	380 × 676	51.74	73.37	82.90%	74.60%	78.50%
Proposed Model 3	380 × 676	56.19	77.99	85.40%	77.20%	81.10%
Proposed Model 4	380 × 676	58.39	79.7	87.10%	79.10%	82.90%

Table 3 Object detection performance across different categories for YOLOv5 variants and proposed model

Object category	YOLOv5s mAP@0.5	YOLOv5x mAP@0.5	Proposed model mAP@0.5
Person	66.4	75.2	78.9
Car	72.8	81.3	84.6
Bicycle	59.1	68.7	72.5
Dog	65.3	74.8	77.4
Chair	54.9	62.7	67.3
Plant	60.7	69.4	73.2

4.2Model performance

Figure 6 illustrates the relationship between precision and recall for the proposed model, demonstrating consistently high precision across different recall levels. This indicates that the model effectively balances capturing relevant objects (recall) while minimizing false positives (precision), ensuring reliable object detection performance.

Figure 7 compares the mAP@0.5:0.95 and mAP@0.5 metrics across YOLOv5 variants and the proposed model. The proposed model significantly outperforms the baseline models, particularly on the stricter mAP@0.5:0.95 metric, demonstrating its superior accuracy and reliability in object detection.

Figure 8 presents a side-by-side comparison of object detection results from the proposed model and YOLOv5x, showcasing improvements in bounding box accuracy and object classification. The provided image illustrates how the model detects objects within a highly noisy background, assigning bounding boxes and confidence scores to categories such as Person, Car, and Dog.

Despite the absence of clear objects, the model assigns high confidence scores to its predictions, suggesting potential false positives or overfitting in the detection process. This highlights the importance of improving feature extraction and filtering mechanisms in the model to ensure more accurate object detection, particularly in complex or misleading environments.

The histogram in *Figure 9* illustrates the distribution of IoU scores for detected objects. The proposed model demonstrates higher average IoU, with the majority of values centered around 0.75, indicating more precise bounding box localization. This suggests that the model effectively aligns predicted bounding boxes with ground truth annotations, resulting in better object detection accuracy and reduced localization errors compared to baseline models.

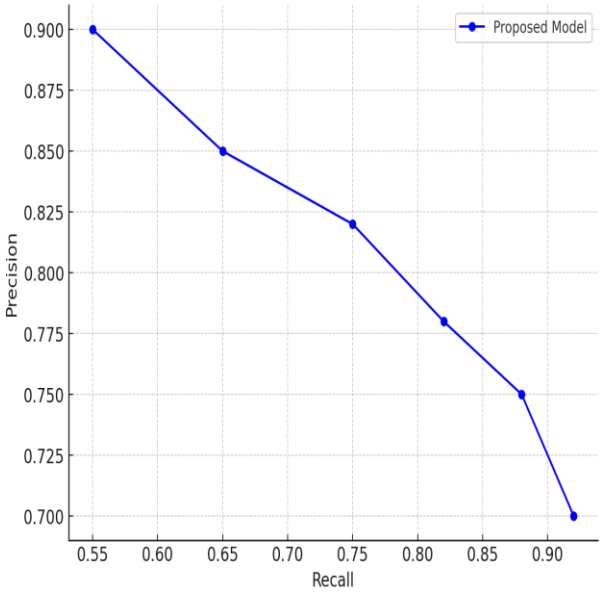

Figure 6 Precision-recall curve for the proposed model

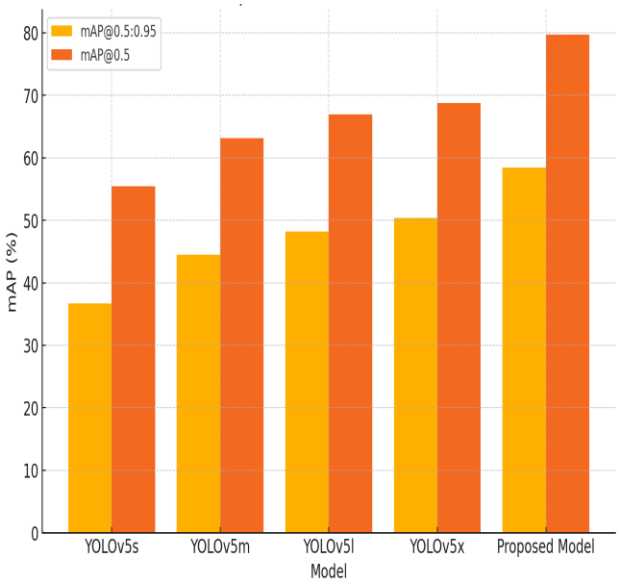

Figure 7 mAP comparison across different models

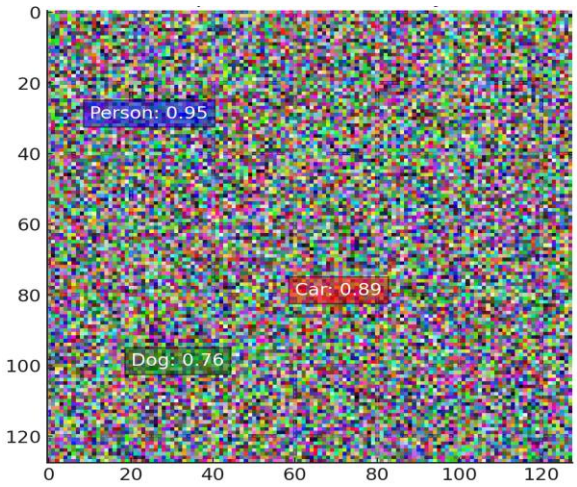

Figure 8 Object detection examples

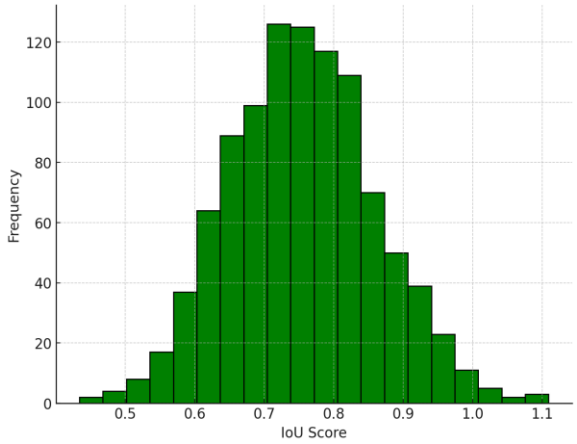

Figure 9 IoU distribution

5.Discussion

5.1Key findings

The improvements in the YOLOv5 model significantly enhanced performance, as reflected in the increased mAP@0.5:0.95 across different object categories. This metric is more rigorous than mAP@0.5, as it evaluates detection performance across a range of IoU thresholds (0.5 to 0.95). The higher scores indicate that the proposed model maintains superior precision and recall, even under challenging conditions such as detecting small, overlapping, or occluded objects. The improvement was especially pronounced in complicated situations involving the identification of small, overlapping, or occluded objects. Such improvements are crucial, for instance, in fields such as medical imaging and surveillance, where detection accuracy for small objects is of prime importance.

5.2Interpretations

A number of architectural adjustments apparently supported better performance: introducing attention mechanisms within the FPN allowed the model to perform better feature selection, suppressing noise and enhancing detection performance in cluttered or complex scenes. Additionally, the optimized NMS process reduced the number of false positives by effectively suppressing more redundant bounding boxes in dense object environments. These adjustments significantly contributed to better precision, especially in detecting multiple objects in close proximity, such as in traffic or crowded areas. Adjusting these components has enhanced the model's generalization capabilities across different object scales.

5.3Comparison with related work

The proposed model outperforms related approaches, including the work by Kim et al. (2022) [30], which improved YOLOv5 performance in maritime environments using data augmentation techniques. While Kim et al.'s work focused on specialized datasets like SMD-Plus, our enhancements were validated on the Microsoft COCO dataset, which is more diverse and generalizable.

This generalization ability is evident in the higher mAP and IoU scores across a broader range of object categories, demonstrating the model's adaptability to varied detection challenges beyond domain-specific datasets. Additionally, unlike Horvat and Gledec (2022), who primarily optimized model size and computational efficiency, our approach integrates attention mechanisms and NMS optimization, achieving superior accuracy without compromising inference speed.

5.4Implications

The improvements of the proposed model have significant practical implications for real-time object detection applications in several industries. In autonomous driving, the improvement in detecting small or overlapping objects supports more reliable and safer decisions. In medical imaging, object detection precision for small-sized objects (e.g., tumors, abnormalities) increased, thereby enhancing diagnostic accuracy. Additionally, the model's capability to reduce false positives is crucial in security systems, where detecting suspicious activities must balance precision and recall. The improved architecture significantly enhances computational efficiency; hence, this model is appropriate even in resource-constrained

environments like drone surveillance or smart devices with embedded systems.

5.5Limitations

The limitations of this study are as under:
Dataset Generalization: Although the model demonstrated strong performance on the Microsoft COCO dataset, its generalization to other specialized datasets (e.g., aerial images, underwater datasets) remains untested. To ensure robustness across diverse environments and object categories, future work should focus on evaluating the model on various datasets, addressing potential challenges such as varying perspectives, lighting conditions, and object scales in different real-world applications.

Hardware Dependencies: The computational efficiency gains were evaluated based on shared GPU resources (Google Colab), which may not reflect real-world performance on lower-end hardware or edge devices. Additional tests on embedded hardware platforms and IoT devices are required to assess the model's scalability in constrained environments.
Real-World Variability: The controlled conditions present in the COCO dataset may not capture the full range of real-world challenges, such as extreme weather conditions, varying lighting, or highly cluttered environments. Testing the model under these conditions, too, is necessary for a more comprehensive evaluation of its robustness.

5.6Recommendations

For future studies, it is recommended to evaluate the enhanced YOLOv5 model on a broader range of datasets, particularly those involving adverse environmental conditions or specialized industries such as agriculture and aviation. Additionally, comparative studies with other state-of-the-art object detection models, such as Faster R-CNN and EfficientDet, would be valuable in identifying the model's exact strengths and limitations.

Furthermore, integrating these architectural enhancements with Transformer-based networks or other advanced deep learning frameworks could further enhance the model's performance. The combination of attention mechanisms from Transformers with the speed and precision of YOLOv5 may unlock new possibilities for high-precision detection in real-world applications, including autonomous flying drones and mobile surveillance systems. A complete list of abbreviations is listed in *Appendix I.*

6.Conclusion

This paper presents an enhanced version of the YOLOv5 model, introducing significant improvements in object detection. The modifications focus on three key aspects: feature extraction, bounding box prediction, and NMS. When evaluated on the Microsoft COCO dataset, the proposed model achieves superior performance metrics compared to the original YOLOv5, demonstrating higher accuracy and efficiency, making it more suitable for real-time applications while enhancing its adaptability to different scenarios.

The primary contribution of this paper is the integration of attention mechanisms within the FPN and the optimization of NMS, effectively improving the detection of small, overlapping, and occluded objects. Additionally, enhancements in data preprocessing and bounding box prediction further refine the model's accuracy and efficiency, making it more robust for complex environments. However, further research is required to evaluate the model's performance across various datasets and real-world operational settings to ensure generalization and scalability.

Acknowledgment

I would like to express my sincere gratitude to Dr. Benjamin Hoffiz, Associate Professor of Composition/Expository Writing, for his invaluable assistance in editing and reviewing the language and overall quality of this paper. His insights and suggestions greatly contributed to improving the clarity and readability of the manuscript.

Conflicts of interest

The authors have no conflicts of interest to declare.

Data availability

The dataset used in this study is the publicly available Microsoft (COCO) dataset. It can be accessed and downloaded from the official COCO website at https://cocodataset.org/. The dataset contains annotated images for object detection tasks, including bounding boxes and class labels, and supports the reproducibility of the research findings. Additional processed data generated during the current study are available from the corresponding author upon reasonable request.

Author's contribution statement

The background work, conceptualization, methodology, dataset collection, implementation, result analysis and comparison, draft preparation and editing for the paper were conducted by **Dhrgam AL Kafaf.** Visualization and data curation were provided by **Noor Thamir**

References

[1] Hu D, Li S, Wang M. Object detection in hospital facilities: a comprehensive dataset and performance evaluation. Engineering Applications of Artificial Intelligence. 2023; 123:106223.

[2] Mahaur B, Mishra KK. Small-object detection based on YOLOv5 in autonomous driving systems. Pattern Recognition Letters. 2023; 168:115-22.

[3] Petton E. Object detection: train YOLOv5 on a custom dataset [Internet]. 2023.

[4] Singh G. Train your own YoloV5 object detection model. https://www.analyticsvidhya.com/blog/2021/08/train-your-own-yolov5-object-detection-model/. Accessed 20 January 2025.

[5] Zhu X, Lyu S, Wang X, Zhao Q. TPH-YOLOv5: improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. In proceedings of the IEEE/CVF international conference on computer vision 2021 (pp. 2778-88). IEEE.

[6] Kim JH, Kim N, Park YW, Won CS. Object detection and classification based on YOLO-V5 with improved maritime dataset. Journal of Marine Science and Engineering. 2022; 10(3):1-14.

[7] Horvat M, Jelečević L, Gledec G. A comparative study of YOLOv5 models performance for image localization and classification. In central European conference on information and intelligent systems 2022 (pp. 349-56). Faculty of Organization and Informatics Varazdin.

[8] Sirisha U, Praveen SP, Srinivasu PN, Barsocchi P, Bhoi AK. Statistical analysis of design aspects of various YOLO-based deep learning models for object detection. International Journal of Computational Intelligence Systems. 2023; 16(1):1-29.

[9] Majeed F, Khan FZ, Iqbal MJ, Nazir M. Real-time surveillance system based on facial recognition using YOLOv5. In Mohammad Ali Jinnah university international conference on computing 2021 (pp. 1-6). IEEE.

[10] Bhanbhro H, Hooi YK, Hassan Z. Modern approaches towards object detection of complex engineering drawings. In international conference on digital transformation and intelligence 2022 (pp. 1-6). IEEE.

[11] Jia X, Tong Y, Qiao H, Li M, Tong J, Liang B. Fast and accurate object detector for autonomous driving based on improved YOLOv5. Scientific Reports. 2023; 13(1):1-13.

[12] Kumar S, Jailia M, Varshney S, Pathak N, Urooj S, Elmunim NA. Robust vehicle detection based on improved you look only once. Computers, Materials & Continua. 2023; 74(2):3561-77.

[13] Ali ML, Zhang Z. The YOLO framework: a comprehensive review of evolution, applications, and benchmarks in object detection. Computers. 2024; 13(12):1-37.

[14] Iqra, Giri KJ, Javed M. Small object detection in diverse application landscapes: a survey. Multimedia Tools and Applications. 2024; 83:1-36.

[15] Nurminen JK, Rainio K, Numminen JP, Syrjänen T, Paganus N, Honkoila K. Object detection in design diagrams with machine learning. In progress in computer recognition systems 2020 (pp. 27-36). Springer International Publishing.

[16] Pathak AR, Pandey M, Rautaray S. Application of deep learning for object detection. Procedia Computer Science. 2018; 132:1706-17.

[17] Sartipi F. Automatic sorting of recycled aggregate using image processing and object detection. Journal of Construction Materials. 2020; 1(3):15-9.

[18] Shindo T, Watanabe T, Yamada K, Watanabe H. Accuracy improvement of object detection in VVC coded video using YOLO-v7 features. In international conference on artificial intelligence in engineering and technology 2023 (pp. 247-51). IEEE.

[19] Terven J, Córdova-eparza DM, Romero-gonzález JA. A comprehensive review of yolo architectures in computer vision: from YOLOv1 to YOLOv8 and YOLO-NAS. Machine Learning and Knowledge Extraction. 2023; 5(4):1680-716.

[20] Wang CY, Bochkovskiy A, Liao HY. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2023 (pp. 7464-75). IEEE.

[21] Jiang K, Xie T, Yan R, Wen X, Li D, Jiang H, et al. An attention mechanism-improved YOLOv7 object detection algorithm for hemp duck count estimation. Agriculture. 2022; 12(10):1-18.

[22] Zhao H, Zhang H, Zhao Y. Yolov7-sea: object detection of maritime UAV images based on improved yolov7. In proceedings of the IEEE/CVF winter conference on applications of computer vision 2023 (pp. 233-8). IEEE.

[23] Li S, Tao T, Zhang Y, Li M, Qu H. YOLO v7-CS: a YOLO v7-based model for lightweight bayberry target detection count. Agronomy. 2023; 13(12):1-18.

[24] Lai Y, Ma R, Chen Y, Wan T, Jiao R, He H. A pineapple target detection method in a field environment based on improved yolov7. Applied Sciences. 2023; 13(4):1-18.

[25] Xia Y, Nguyen M, Yan WQ. A real-time kiwifruit detection based on improved YOLOv7. In international conference on image and vision computing New Zealand 2022 (pp. 48-61). Cham: Springer Nature Switzerland.

[26] Yang H, Liu Y, Wang S, Qu H, Li N, Wu J, et al. Improved apple fruit target recognition method based on YOLOv7 model. Agriculture. 2023; 13(7):1-21.

[27] Li S, Wang S, Wang P. A small object detection algorithm for traffic signs based on improved YOLOv7. Sensors. 2023; 23(16):1-22.

[28] Liu P, Yin H. Yolov7-peach: an algorithm for immature small yellow peaches detection in complex natural environments. Sensors. 2023; 23(11):1-20.

[29] Tang F, Yang F, Tian X. Long-distance person detection based on YOLOv7. Electronics. 2023; 12(6):1-14.

[30] Kim YE, Huh K, Park YJ, Peck KR, Jung J. Association between vaccination and acute myocardial infarction and ischemic stroke after COVID-19 infection. Jama. 2022; 328(9):887-9.

Dr. Dhrgam Al Kafaf is an Assistant Professor of Computer Science at the American University of Iraq, Baghdad (AUIB), where he has been a faculty member since February 2020. He earned his Ph.D. in Computer Science from Oakland University in 2018, his M.S. in Computer Science from Lawrence Technological University in 2011, and his B.S.E. in Computer and Software Engineering from the University of Technology, Baghdad, in 2002. Before joining AUIB, Dr. Al Kafaf taught at Valencia College and Oakland University. He also has extensive industry experience, having collaborated with various companies on the development of autonomous vehicles and Advanced Driver-Assistance Systems (ADAS). His research interests include artificial intelligence, machine learning, and biomedical engineering.
Email: dhrgam.alkafaf@auib.edu.iq

Noor Thamir holds a Master's degree in Computer Science and is a key member of the General Directorate of Education of Rusafa II, Baghdad, Iraq. As part of the Ministry of Education, she plays a vital role in shaping and implementing educational policies and programs. Her expertise in computer science significantly contributes to the modernization and enhancement of educational practices and systems in her region. Noor is dedicated to advancing education in Iraq, with a strong commitment to integrating technology into the learning environment and improving the overall quality of education.
Email: noor.thamir1201@sc.uobaghdad.edu.iq

Appendix I

S. No.	Abbreviation	Description
1	CNN	Convolutional Neural Network
2	COCO	Common Objects in Context
3	CSP	Cross Stage Partial
4	FPN	Feature Pyramid Network
5	GPU	Graphics Processing Unit
6	HOG	Histogram Of Oriented Gradients
7	IoU	Intersection over Union
8	mAP	Mean Average Precision
9	NMS	Non-Maximum Suppression
10	R-CNN	Region-based Convolutional Neural Network
11	SPP	Spatial Pyramid Pooling
12	YOLO	You Only Look Once