Frontiers in Engineering and Informatics
ISSN: 3049-3412 (Online)
Volume No: 01 – Issue No: 03 – 10 September 2025

A Hybrid Deep Learning Approach for Accurate and Efficient Object Detection

Perumalla Naga Padmavathi[1*]

[1*]Department of Computer science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India.

Corresponding Author E-mail ID: [1]padmavathi481@gmail.com

Abstract

Humans can easily identify multiple objects in an image or video but for computers, it is very difficult to identify. The procedure of precisely detecting and promptly recognizing items is quite challenging. However, with the assistance of diverse object detection algorithms, we can now do this work with utmost precision. The proposed method achieved an accuracy of 0.78 to 0.84 on various objects. The advantage of embedding mask RCNN and yolo v7 was achieved with a good precision value. The experimental results were published by masking the specific object and masking the background of the image. The Advantage of using Embedded with Mask R-CNN and YOLO v7 is that it achieves a good precision value. The experimental results were published with masking applied to a specific object and its background. The proposed method concluded that, with YOLO v7, we could reduce the computational effort by 30% and parameter optimization by 40% compared to the existing method.

Keywords: Object detection, Deep Learning, YOLO, CNN

1. INTRODUCTION

An interdisciplinary field called computer vision focuses on giving robots the ability to comprehend and interpret the visual world through digital photos or movies. It entails the creation of algorithms and methods that enable computers to extract relevant data from digital photos, videos, or other visual inputs. Computer vision aims to enable robots to interpret, evaluate, and comprehend the visual environment like humans Brownlee J et al, (2019) Computer vision has a wide variety of uses, from straight forward tasks like face recognition and object identification to more involved ones like autonomous cars, medical picture analysis, and augmented reality. Several industries employ computer vision, including manufacturing, robotics, entertainment, security, and healthcare. Computer vision algorithms frequently employ methods from various disciplines, including artificial intelligence, machine learning, computer graphics, and signal processing, to accomplish these tasks. Recent developments in deep learning have significantly transformed computer vision, enabling robots to perform remarkably well across a range of visual identification tasks Xiao Y et al ,(2020).

Deep learning, a subfield of machine learning Bai Q et al,(2020), focuses on training artificial neural

DoI: https://doi.org/10.63949/crinfo.v1i3.003

networks to identify and interpret visual and auditory stimuli, analyze natural language, and make informed judgments. The field of artificial intelligence has undergone a significant transformation with the advent of deep learning techniques. This advancement has enabled computers to acquire knowledge and enhance their performance in tasks that were previously challenging or unfeasible to automate. Deep learning algorithms are built upon artificial neural networks, which aim to mimic the structure and

functioning of the human brain. These neural networks comprise neurons or interconnected nodes that process and transmit data through weighted connections Talukdar J et al, (2018). During the training process, the weights assigned to these connections are adjusted to optimize the network's effectiveness for a specific task. Joseph Redmon, Ali Farhadi, and other researchers at the University of Washington developed a state-of-the-art real-time object identification system known as You Only Look Once (YOLO). Following the application of a convolutional neural network to an input image, the method proceeds to generate predictions for bounding boxes, objectless scores, and class probabilities for every individual cell inside the grid Vijayakumar A et al, (2024).

The YOLO algorithm Zhou Y et al, (2024)uses a single neural network to make predictions, as opposed to other object detection algorithms that use multiple networks in a pipeline. This makes YOLO faster and more efficient than other methods, allowing it to achieve real-time performance on a wide range of hardware platforms. YOLO is available in many iterations, with YOLOv5 being the latest iteration. The extensive utilization of this technology in various domains, including image and video analysis, surveillance systems, and autonomous vehicles, can be attributed to its exceptional performance on multiple object detection benchmarks.

The concept of YOLO has gained particular popularity in recent years, especially among younger generations seeking to break free from societal expectations and embrace their individuality. This can manifest in various ways, such as traveling to new places, trying extreme sports, pursuing unconventional career paths, or even just speaking up for oneself and taking risks in interpersonal relationships. However, it's essential to note that the idea behind YOLO can also be taken too far, and it's crucial to strike a balance between the pursuit of adventure and responsibility and common sense. For example, taking unnecessary risks or engaging in reckless behavior can have serious consequences, both for oneself and for others Kang S et al, (2025).

The objective of this paper is to develop an image processing system that can accurately detect suspicious objects in an image using a combination of two deep learning algorithms: Mask R-CNN and YOLOv7. The Mask R-CNN algorithm is a state-of-the-art object detection and segmentation model that can accurately segment the object of interest in an image. This algorithm utilizes a combination of a region proposal network (RPN) and a Fully Convolutional Network (FCN) to generate object proposals and segment the object of interest, respectively. Mask R-CNN is particularly useful in detecting objects

DoI: https://doi.org/10.63949/crinfo.v1i3.003

that are occluded or have complex shapes.

The YOLOv7 algorithm, on the other hand, is a real-time object detection algorithm that can detect objects in an image at high speeds. YOLOv7 uses a single convolutional neural network (CNN) to perform object detection and classification in a single step. It is known for its fast processing speed and can detect objects in real-time applications. It will be measured based on the accuracy and speed of detection of the suspicious objects in the input image. It is expected that the combination of Mask R-CNN and YOLOv7 will improve the accuracy of object detection by localizing and isolating the object of interest, while also providing fast processing speeds for real-time applications.

2. RELATED WORK

Region-based convolutional neural networks have improved object detection over earlier techniques, such as HOG and SIFT Girshick R et al, (2015). Selective features, which typically range from 2000 to 20000 features, are used in R-CNN models to extract the most salient traits. The process of identifying the most significant extractions can be computed using a selective search method, which yields more insightful regional recommendations. The input of the convolutional neural network is a 224 x 224 fixed-size RGB image. Eliminating the average RGB values corresponding to every pixel in the training dataset is the main preprocessing step. The image is next subjected to a sequence of convolutional (Conv.) layers, each of which contains filters with a somewhat small 3 3 receptive field. The smallest receptive field size required to adequately represent the ideas of left/right, up/down, and center regions is this one.

Fast R-CNN combines multiple innovations to improve detection accuracy, which speeds up the training and testing procedure Cao C et al, (2019). Fast R-CNN trains the highly deep VGG16 network nine times quicker than R-CNN and surpasses R-CNN in terms of mean average precision (mAP), according to PASCAL VOC 2012. One of the most remarkable revisions of R-CNN is the quicker R-CNN model, which performs noticeably quicker than its earlier incarnations. To generate region proposals, the R-CNN and Fast R-CNN models use a selective search method. However, unlike the current approach, the Faster R-CNN technique utilizes an enhanced region proposal network. To generate useful outputs, the Region Proposal Network (RPN) computes images at various scales Steno P et al, (2021). Due to the increasing popularity of this architectural style, Residual Blocks were developed as a solution for vanishing or developing slopes. The network employs a methodology known as skip connections. The skip connection traverses multiple levels to establish connections between layer

DoI: https://doi.org/10.63949/crinfo.v1i3.003

activations and subsequent layers. Hence, a residual block is generated. The remaining blocks are arranged in a stack to form reservoirs.

The system's architecture Zheng Y et al, (2011) integrates lateral connections, a top-down pathway, and a bottom-up pathway to establish connections between low-resolution and high-resolution components. The bottom-up approach allows for the utilization of any image as input, irrespective of its dimensions. The bottom-up method imposes no limitations on the dimensions of the input image. Following the processing of data by convolutional layers, pooling layers are employed to reduce the dimensionality of the data. It is important to acknowledge that each collection of feature maps of identical size is commonly known as a stage, and the features employed for the pyramid level are derived from the conclusion layer in each stage.

The feature maps from the appropriate level of the top-down pathway are included into the final stage of the bottom-up pathway through lateral connections. The top-down method employs unspooling to increase the size of the final feature maps. The top-down methodology employs lateral connections to undergo unspooling, thereby upsampling the final feature maps. Furthermore, the feature maps are enhanced through the incorporation of feature maps derived from the bottom-up pathway that are situated at an equivalent level. The feature maps of the pathway are combined and examined using a 1x1 matrix in a bottom-up manner to establish the connections Choi HT et al, (2021).

DenseNet Hou Y et al, (2024) utilizes feedforward connections to establish connections between each layer and all other layers. The network we have developed has L(L+1)/2 direct connections, in contrast to the L connections observed in traditional convolutional networks with L layers. In conventional convolutional networks, each layer is connected to the subsequent layer through a single link. Each subsequent layer utilizes the feature maps from preceding layers as inputs, whereas each subsequent layer employs its own feature maps as inputs. DenseNets offer several notable advantages, including addressing the problem of vanishing gradients, enhancing feature propagation, facilitating feature reuse, and substantially reducing the number of parameters Kumar P et al, (2024).

The Mask R-CNN is a Faster R-CNN model that consists of three output branches Weng X et al, (2025). It initially computes the bounding box coordinates, followed by the appropriate class, and last generates the binary mask3 to segment the object. An FCN generates a binary mask for a certain region of interest (ROI) with a predetermined size. To mitigate misalignments caused by the quantization of the RoI coordinates, a RoIAlign layer is utilized instead of a RoIPool. The Mask R-CNN model stands out due to its multi-task loss, which encompasses the losses of the segmentation mask, the predicted class, and the bounding box coordinates. The strategy aims to address mutually beneficial tasks, enhancing performance on each task individually.

DoI: https://doi.org/10.63949/crinfo.v1i3.003

3. METHODOLOGY

The aim of this work is to develop an image processing system that can accurately and efficiently detect suspicious objects in an image by using a combination of two advanced deep learning algorithms: Mask R-CNN and YOLOv7. The system will be designed to take an input image, segment and isolate the object of interest using Mask R-CNN, and then perform object detection and classification on the masked object using YOLOv7.

The existing method of integrating Mask R-CNN with YOLOv5 involves using the feature maps generated by Mask R-CNN as input to the YOLOv5 detection network. The feature maps are first extracted from the region of interest (ROI) proposals generated by Mask R-CNN. The ROI proposals are generated using a region proposal network (RPN) which suggests potential object locations in the image.

The YOLOv5 detection network then processes the feature maps extracted from the ROI proposals to detect objects. The detection network consists of a series of convolutional layers that progressively refine the feature maps to predict bounding boxes, objectness scores, and class probabilities. The YOLOv5 detection network is designed to be lightweight and fast, using a series of novel architectural and optimization techniques.

The integration of Mask R-CNN with YOLOv5 allows the model to leverage the strengths of both methods shown in Figure 1 and Figure 2. Mask R-CNN is a state-of-the-art object detection algorithm that can accurately segment objects in an image, while YOLOv5 is a lightweight algorithm that can detect objects quickly. The combination of these two algorithms allows for the accurate and fast detection of objects in an image. This approach has been shown to detect objects with an accuracy of 0.85 to 0.91, which is relatively high. However, the computational power required by this method can be a significant drawback in real-world applications. This is particularly true for applications where real-time detection is required, such as in autonomous vehicles or robotics.

To address the issue of speed and high computational power, several approaches can be taken. One approach is to optimize the integration process by reducing the number of feature maps generated by Mask R-CNN, thereby reducing the overall system's computational cost. Another approach is to modify the architecture of YOLOv5 to make it more efficient in processing the feature maps generated by Mask R-CNN.

DoI: https://doi.org/10.63949/crinfo.v1i3.003

Frontiers in Engineering and Informatics
ISSN: 3049-3412 (Online)
Volume No: 01 – Issue No: 03 – 10 September 2025

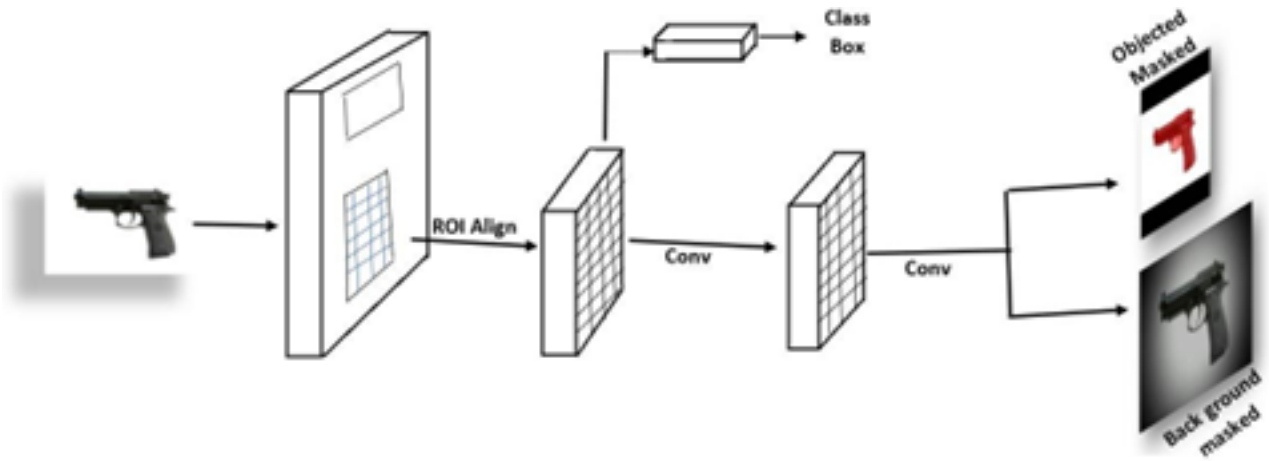

Figure 1 Mask RCNN Architecture

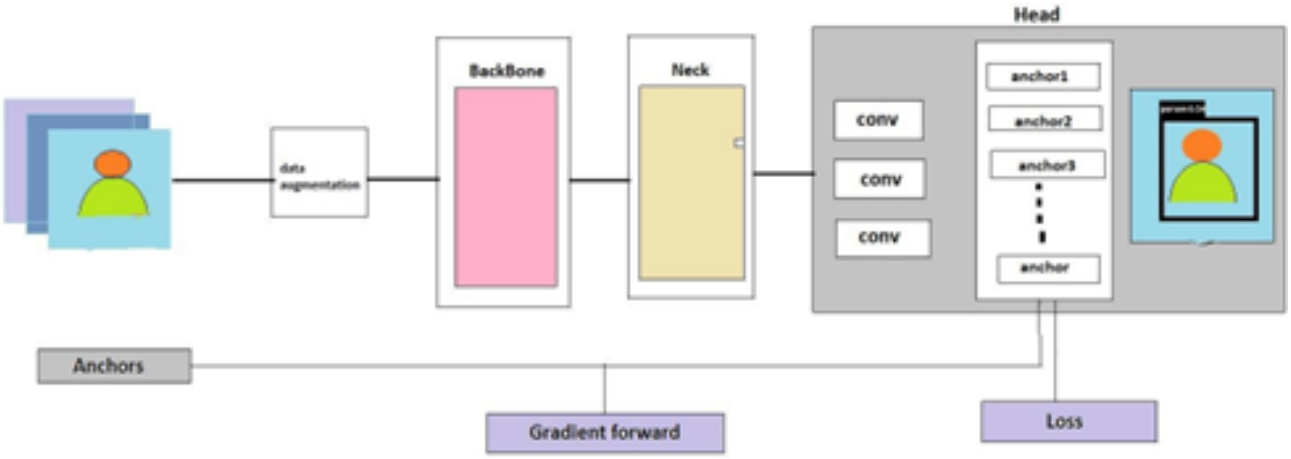

Figure 2. Working of YOLO V7

To improve the speed and computational efficiency of the existing method that integrates Mask RCNN with YOLOv5, one proposed approach could be to use YOLOv7 with Mask R-CNN. YOLO v7 can reduce 40% of parameters in the image as well as reduces 30% computational power also. while implementing only with YOLO v7 some features are missed to overcome those problems we use mask RCNN to be integrated into the YOLO v7. To implement this proposed method, the YOLOv7

DoI: https://doi.org/10.63949/crinfo.v1i3.003

architecture would need to be modified to accept feature maps from a masked image as input instead of the original image. The modified YOLOv7 would then perform object detection on the feature maps, which would significantly reduce the computational cost of the overall system.

Mask R-CNN is a cutting-edge Convolutional Neural Network that demonstrates exceptional performance in the domains of instance and picture segmentation. The Mask R-CNN model was based on the Faster R-CNN architecture, which is a Region-Based Convolutional Neural Network known for its great effectiveness. To have a comprehensive understanding of the functioning of Mask R-CNN, it is important to initially grasp the fundamental principle of image segmentation.

The task of computer vision, the process of image segmentation involves the division of a digital image into many segments, which are comprised of individual pixels referred to as image objects. Segmentation is a technique employed to precisely determine the position and delineate boundaries and entities, such as curves and lines.

Mask R-CNN, alternatively referred to as Mask RCNN, is a highly sophisticated Convolutional Neural Network (CNN) that demonstrates exceptional performance in applications such as image segmentation. The development of Mask R-CNN was built on the underlying design of Faster R-CNN, which is a region-based convolutional neural network. Comprehending the concept of picture segmentation is essential for comprehending the functioning of Mask R-CNN.

The objective of computer vision The process of image segmentation involves the partitioning of a digital image into many segments, wherein pixels are divided into distinct groupings referred to as image objects. The segmentation process involves the identification and delineation of boundaries and characteristics, such as curves and lines.

Mask R-CNN employs a convolutional neural network (CNN) as its fundamental architecture to extract features from the input image. The backbone network often consists of a pre-trained network, such as ResNet or ResNeXt, that has been trained using a large dataset like ImageNet.

RPN works by sliding a small window (called anchor) over the feature map of the input image and predicting the probability of each anchor being an object or not. In other words, for each anchor, RPN predicts a binary label indicating whether the anchor contains an object or not, as well as four additional values that represent the offset of the anchor with respect to a ground-truth bounding box. These predicted values are used to adjust the anchor's position and size to better fit the object.

ROI Align layer differs from the previously used ROI Pooling layer in the way it computes the feature maps for each region. In the ROI Pooling layer, each region is divided into a fixed number of equal-sized sub-windows, and the feature map for each sub-window is computed using max pooling.

After acquiring the individual ROI feature map, it is possible to make predictions regarding the object

DoI: https://doi.org/10.63949/crinfo.v1i3.003

category and generate a more accurate instance bounding box. The aforementioned branch is a fully connected layer that establishes the bounding box coordinates for 4n instances and a mapping between feature vectors and the ultimate n classes. The Mask Generation Head takes the feature map for each proposed region (obtained using the ROI Align layer) and applies a small convolutional neural network (CNN) to predict a binary mask for the object within the region. The CNN typically consists of a series of convolutional layers followed by a final deconvolutional layer, which up samples the feature map to the original size of the proposed region.

The output of the Mask Generation Head is a binary mask with the same size as the proposed region. Each pixel in the mask is assigned a value of 1 if it belongs to the object within the region and 0 otherwise. The predicted mask is then used to segment the object in the input image.

The Extended Efficient Layer Aggregation Network (EELAN) is designed to address the trade-off between model efficiency and accuracy in deep neural networks, particularly for large-scale image recognition tasks. It builds upon the Efficient Layer Aggregation Network (ELAN) architecture, a lightweight and efficient framework for image classification tasks. EELAN extends ELAN by introducing additional blocks to enhance the model's representation capability. These blocks include the Global Context Block (GCB) for capturing global contextual information, the Dual Attention Module (DAM) for enhancing feature representation, and the Multi-Path Refinement Module (MPRM) for improving the model's robustness to occlusions and image corruptions.

EELAN demonstrates exceptional performance on many benchmark image classification datasets, surpassing the bulk of existing deep learning architectures in terms of efficiency. The primary function of the neck is to generate feature pyramids. In the domain of object scaling, the utilization of feature pyramids facilitates the enhanced generalization of models. The ability to identify an object across many scales and sizes is advantageous. Feature pyramids are highly advantageous and enhance the performance of models when employed with unfamiliar data. The primary responsibility of the model's head is to execute the ultimate detection phase. The proposed methodology involved applying anchor boxes to the features, resulting in the generation of final output vectors that included bounding boxes, class probabilities, and objectness scores.

DoI: https://doi.org/10.63949/crinfo.v1i3.003

Frontiers in Engineering and Informatics
ISSN: 3049-3412 (Online)
Volume No: 01 – Issue No: 03 – 10 September 2025

Algorithm 1 - Hybrid Mask R-CNN–YOLOv7 Framework for Suspicious Object Detection and Classification

Pseudocode

Input: Images with suspicious objects

Output: suspicious objects can be classified by the bounding boxes with accuracy and with classname

Pseudocode:

1) Load the image into memory

2) Apply Mask R-CNN to identify regions of interest (ROI) in the image

3) Load the Mask R-CNN model into memory

4) Preprocess the image (resize, normalization, etc.)

5) Pass the preprocessed image through the Mask R-CNN model to obtain the output

6) Postprocess the output to obtain the list of ROIs and their corresponding masks 3. For each ROI identified by Mask R-CNN:

7) Extract the masked object

8) Apply the mask to the original image to obtain the masked object ii. Convert the masked object to a format compatible with YOLOv7 (e.g., resize, normalization)

9) Pass the masked object through YOLOv7 to detect suspicious objects

10) Load the YOLOv7 model into memory ii. Pass the masked object through the YOLOv7 model to obtain the output iii. Postprocess the output to obtain the list of suspicious objects and their corresponding confidence scores

11) If any suspicious objects are found, mark the object as suspicious by:

12) Drawing a bounding box around the object on the original image ii. Labeling the object with the corresponding class and confidence score

13) Output the original image with any suspicious objects identified

4. RESULTS AND DISCUSSION

Implementation of this work was done using Google Colab. To import data into Colab, we must upload the data to the drive and then import it by mounting the drive. The different libraries, such as TensorFlow and Keras, will be imported into Colab. After importing the data and libraries, I mask the

DoI: https://doi.org/10.63949/crinfo.v1i3.003

Frontiers in Engineering and Informatics
ISSN: 3049-3412 (Online)
Volume No: 01 – Issue No: 03 – 10 September 2025

background of the images, and four outputs are displayed for the user. By using the save fig method, the output images are downloaded into the runtime. The Figure 3 describes the sample images for the object detection

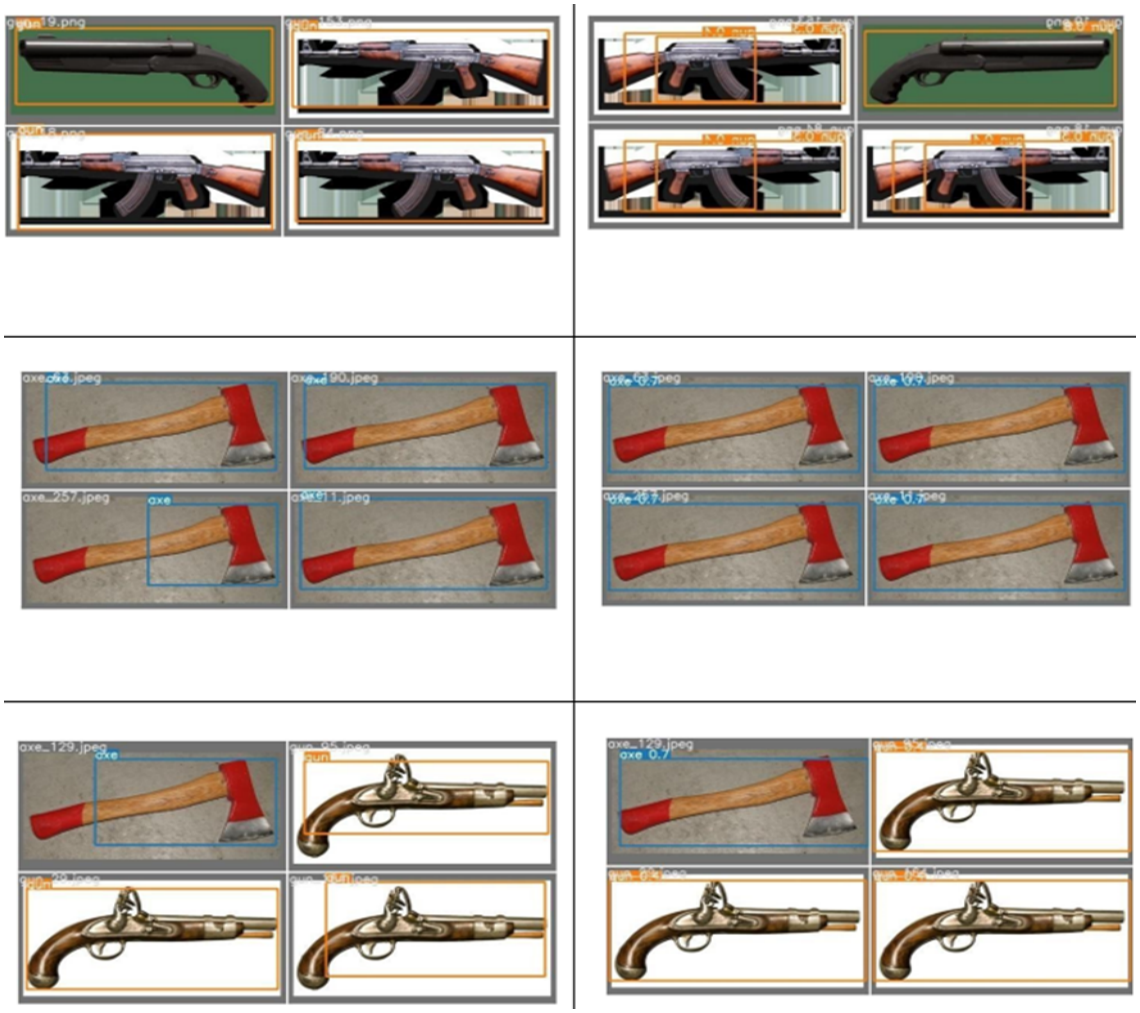

Figure 3: Testing results

DoI: https://doi.org/10.63949/crinfo.v1i3.003

Frontiers in Engineering and Informatics
ISSN: 3049-3412 (Online)
Volume No: 01 – Issue No: 03 – 10 September 2025

Table 1: Performance Analysis

Image	Yolo V7	Yolo v7 & MRCNN (Object was Masked)	Yolo v7 & MRCNN (Background was Masked)	Analysis
	0.71	0.74	0.75	Yolov7 with background masked image has high accuracy
	0.83	0.75	0.84	Yolov7 with background masked image has high accuracy
	0.51	0.51	0.51	Accuracy remains constant
	0.61	0.58	0.60	Yolov7 and background masked image have closed accuracy

DoI: https://doi.org/10.63949/crinfo.v1i3.003

Frontiers in Engineering and Informatics
ISSN: 3049-3412 (Online)
Volume No: 01 – Issue No: 03 – 10 September 2025

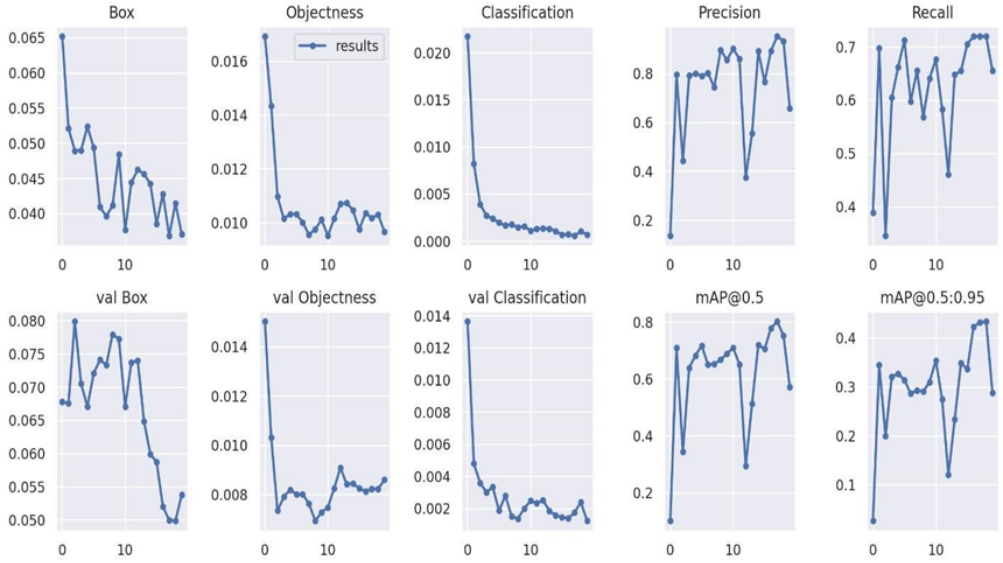

Figure 4. Testing Results

Table 1 describes the performance of the proposed mode. The box loss, object loss, cls loss are used to show the predicted bounding box, object presence and the correctness of the classification. Figure 4 discuss about the class loss, box loss and the object class while doing the train and validation of the project in yolo-v5. 1.while doing this training the box loss decreased from 0.06 to 0.03 and the object loss decreased from 0.016 to 0.010 and the class loss was decreased from 0.02 to 0.00.while doing this validation the box loss decreased from 0.08 to 0.055 and the object loss decreased from 0.0105 to 0.01 and the class loss was decreased from 0.014 to 0.002.The precision and recall are used for the calculation of the positive samples and these precision was increased from the 0.0 to 0.65 and the recall was increased from 0.35 to 0.65.The mean average precision was increased from the 0.0 to 0.6.

5. CONCLUSION

The field of object-detecting technologies has been developing for decades. The object-detecting effect and method have advanced from conventional computer vision technology to today's well- liked deep learning technology. Nevertheless, the more the effect and effectiveness of the algorithm, the greater the network and hardware needs, which is the application of the method is subject to a fair amount of limitation. Furthermore, the present deep learning-based detection approach requires enough training data to build the model and produce a reliable detection result. By using this dissertation and based on experimental results we are able to detect axe, guns, knives, and swords with high speed and good precision when we apply a mask to the background of the image, and it can’t identify the two different

DoI: https://doi.org/10.63949/crinfo.v1i3.003

classes of the object in the picture. Without masking also it can identify the objects but with less accuracy when compared to the masked image, when we worked with Yolo v5 it detects the object with high accuracy, but with yolo v7 it reduces 40% parameters for increasing the computational power that's why if we masked the image and then give that masked image to the model it will gives better results. Yolov7 is still a young algorithm that is still being developed. The problems addressed by developers have a lot of space for improvement. if we modify our proposed method with the feature maps generated by the mask RCNN is directly given as input to the Yolo v7. The method will be very helpful in resolving many computer vision problems once it becomes widely used.

REFERENCES

[1]. Brownlee, J. (2019). *Deep learning for computer vision: Image classification, object detection, and face recognition in Python*. Machine Learning Mastery.

[2]. Xiao, Y., Tian, Z., Yu, J., Zhang, Y., Liu, S., Du, S., & Lan, X. (2020). A review of object detection based on deep learning. *Multimedia Tools and Applications*, 79(33), 23729–23791.

[3]. Bai, Q., Li, S., Yang, J., Song, Q., Li, Z., & Zhang, X. (2020). Object detection recognition and robot grasping based on machine learning: A survey. *IEEE Access*, 8, 181855–181879.

[4]. Talukdar, J., Gupta, S., Rajpura, P. S., & Hegde, R. S. (2018). Transfer learning for object detection using state-of-the-art deep neural networks. In *Proceedings of the 5th International Conference on Signal Processing and Integrated Networks (SPIN)* (pp. 78–83). IEEE.

[5]. Vijayakumar, A., & Vairavasundaram, S. (2024). YOLO-based object detection models: A review and its applications. *Multimedia Tools and Applications*, 83(35), 83535–83574.

[6]. Zhou, Y. (2024). A YOLO-NL object detector for real-time detection. *Expert Systems with Applications*, 238, 122256.

[7]. Kang, S., Hu, Z., Liu, L., Zhang, K., & Cao, Z. (2025). Object detection YOLO algorithms and their industrial applications: Overview and comparative analysis. *Electronics*, 14(6), 1104.

[8]. Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2016). Region-based convolutional networks for accurate object detection and segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1), 142–158.

[9]. Cao, C., Wang, B., Zhang, W., Zeng, X., Yan, X., Feng, Z., Liu, Y., & Wu, Z. (2019). An improved Faster R-CNN for small object detection. *IEEE Access*, 7, 106838–106846.

[10]. Steno, P., Alsadoon, A., Prasad, P. W. C., Al-Dala'in, T., & Alsadoon, O. H. (2021). A novel enhanced region proposal network and modified loss function: Threat object detection in secure screening using deep learning. *The Journal of Supercomputing*, 77(4), 3840–3869.

[11]. Zheng, Y., Meng, Y., & Jin, Y. (2011). Object recognition using a bio-inspired neuron model with bottom-up and top-down pathways. *Neurocomputing*, 74(17), 3158–3169.

DoI: https://doi.org/10.63949/crinfo.v1i3.003

[12]. Choi, H. T., Lee, H. J., Kang, H., Yu, S., & Park, H. H. (2021). SSD-EMB: An improved SSD using enhanced feature map block for object detection. *Sensors*, 21(8), 2842.

[13]. Hou, Y., Wu, Z., Cai, X., & Zhu, T. (2024). The application of improved DenseNet algorithm in accurate image recognition. *Scientific Reports*, 14(1), 8645.

[14]. Kumar, P., & Alwakid, G. N. (2024). Improved feature extraction and object detection accuracy with the novel DenseNet algorithm compared to the SqueezeNet algorithm in remote sensing images. In *Proceedings of the IEEE 9th International Conference on Engineering Technologies and Applied Sciences (ICETAS)* (pp. 1–7). IEEE.

[15]. Weng, X., Ma, Q., Li, Q., & Wang, W. (2025). Improved Mask R-CNN algorithm: Multi-ore detection and positioning based multi-sensor fusion in complex field environment. *Measurement*, 246, 116602.

DoI: https://doi.org/10.63949/crinfo.v1i3.003