Object Recognition in Images Using Hybrid Deep Learning Model

Jashanpreet Singh[1], Dr. Rajiv Kumar[2]

[1]Department Of Computer Applications, Rimt University, India
[2]Department Of Computer Applications, Rimt University, India

Abstract: Classifying and identifying objects through images and making bounding boxes is the basicobjective of object recognition and detection.Object recognition, is the most crucial problem,this being the reason it hasreceived a strong attention for the research. With the huge growth of object detection technology in computer vision over the last few years, the subject has seen a significant change. In the 1990s, people were still using creative thought and long-lasting design to figure out how to recognize objects in early computer vision. If you look at how we identify objects today as a change made possible by deep learning, you can learn both high-level and low-level features. This paper discusses blended approach in the field of object recognition through deep learning. Major contribution of this work is to present a hybrid classifier approach with some of prominent backbone architecture using EfficientNet CNN Deep learning model combined with YOLO detector for the object recognition named E-YOLO.On some metrics this model test with some existing model on MS COCO dataset for the Common benchmark. Lastly comparison of the performance and accuracy of existing model with proposed model on these metrics has been discussed. As a result the accuracy of proposed model is better than the existing model.

Keywords: *Object Detection, Deep Learning, Computer Vision, YOLO, EfficientNet*

1. Introduction

Recognizing objects by machines is hard, and people are still faster and better at it than computers. Many algorithms have been developed for object recognition but still, it suffers from complexity and recognition rate. The performance of object detection or recognition algorithm struggles due to many variations in images of objects belonging to the same object categories. The factors likescaling, lighting on object, object's location, viewpoint, and rotation of the objectaffect object recognition.

The efficiency and accuracy of recognition rate depends on the descriptors of the object [1]. It is used to talk about what makes a thing unique. The object recognition algorithm uses many details to figure out what an item is, such as its color, texture, shape, and edges. Shape descriptor is a very useful tool for image pre-processing and it has been widely used for object recognition. An object influenced by distortion and noise in an image can be represented by a high-quality shape descriptor.

The first models for finding objects were made with feature extractors that were put together by hand. These included the Histogram of Oriented Gradients (HOG) model [2] and the Viola-Jones detector [3]. These models performed poorly on unknown datasets and were incredibly weak and inaccurate. With the rise of convolutional neural networks (CNNs) and deep learning for picture recognition, the way we think about finding objects has changed. It was used by Alex Net in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012 [4], which led to more research into how it could be used in computer vision. These days, self-driving automobiles, identity detection, security, and medical applications all use object detection. It has experienced exponential growth in the last several years because to the quick development of new tools and methods.

This paper implements an automatic classification and detection of objects using hybrid approach of deep learning.This kind of approach can be applied to self-driving cars, identity detection, education, industries, supermarkets, and people everywhere to help them learn about various objects for learning, security, and medical purposes.

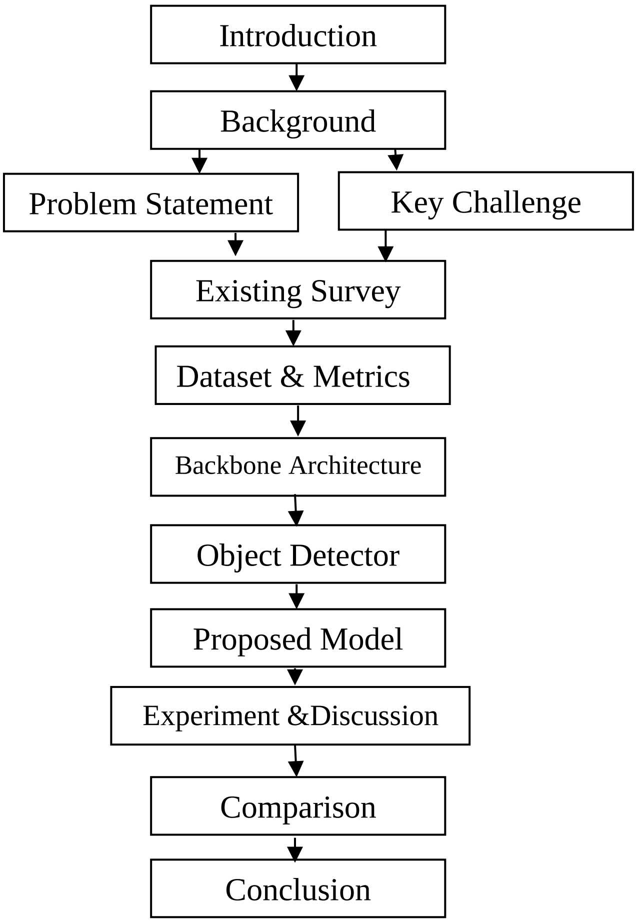


Fig.1. Structure of the paper

In Fig. 1We have discussed systematically object detection Architecture with knowing background problem and also proposed a hybrid model to solve these problems. Firstly, we will read about its introduction in first section. In the second part, we talk about the background of object detection, as well as the problem and the difficulties that come with it. In the third part, we talk about the need for and problems with this survey. Section fourth explains the data set and metric evaluation are listed. Section fifth explains the backbone architecture and object detector used for the proposed model. Section sixth how the proposed model works. In Section seventh result and analysis are discussed.

2. Problem Statement

The main goal of object recognition is to find things in an image. It is basically an extension of object categorization. Identifying every characteristic of the predefined classes is the aim of object detection. It also used the boxes that were lined up along the axes to show where the item that was detected was. For the detector to work, it needs to be able to find all instances of the object classes and draw a square around the item. It is thought to be a difficulty of supervised learning. Modern object detection models are trained on large sets of annotated picture datasets. These models are then tested against a number of accepted standards.

Key challenges in object detection

There are a few significant obstacles in object recognition. Among the main obstacles that deep learning model networks in practical applications must overcome are:

- ✓ *Intra class variation*: Intra class variation between the instances of same object are relatively very common in nature. This variation could be due to numerous reasons like scaling, lighting on object, object's location, viewpoint, and rotation of the object occlusion, illumination etc. Certain objects may be hidden by other things, which makes them harder to get out. These outside factors that aren't limited can have big effects on how a thing looks [5]. It is anticipated that the object picture will be fuzzy, rotated, scaled, or exhibit non-rigid deformation.
- ✓ *Number of categories*: It's hard to solve because there are so many kinds of objects that can be put into groups. The high-quality annotated data is required, which is very difficult to difficult to locate. It is up for debate whether or not to train a detector with fewer examples.
- ✓ *Efficiency*: These days, models need a lot of computing power to get better and more accurate results when they try to find objects. The advancement of computer vision technology depends on the development of effective object detectors. [6]

The goal of this project was to make convolutional neural networks work better in applications that need to recognize objects. Other object recognition methods have made the images that go into the neural network less complicated, which has made CNN even more successful.

3. Literature Review

A lot of reviews of object trackers have come out in the last few years [7–19], as summarized in Table 1. Since then, a lot of new and better deep learning models have come out because object recognition and computer vision in general have changed so quickly.

Archana et al.(2024) Analyzing and interpreting images requires image processing methods such denoising, enhancement, segmentation, feature extraction, and classification. Automated feature extraction is done using Deep Learning (DL) models like Self2Self NN and Denoising CNNs. Though promising, R2R and LE-net image enhancing approaches lack realism. PSPNet and Mask-RCNN segment objects accurately. CNN and HLF-DIP feature extraction automates attribute recognition but is difficult. Residual Networks and CNN-LSTM classifiers are precise yet difficult to understand and compute. This overview emphasises image processing dependability and computational power restrictions to aid decision-making.

Ouf et al.(2023) AI was used to find leguminous seeds for smart farming, which was the main point of the study. It was able to automatically sort and find different kinds of seeds in a variety of settings. The dataset had 828 pictures of leguminous seeds with different backgrounds, shapes, and amounts of people in them. There was a mean average accuracy (mAP) of 98.52% for the machine learning model YOLOv4, which was better than other methods. It also recognized things faster. The study found that YOLOv4 can find leguminous seeds in a number of different situations, making it a useful tool for real-time identification in smart farming apps.

Arkin et al.(2023) Detecting objects is one of the most important problems in computer vision. Convolutional Neural Network (CNN) methods have become common since AlexNet came out. To get better and more accurate detection, researchers looked at a number of different structures. The Transformer, which is well-known in Natural Language Processing, showed potential in computer vision tasks and did better than some CNN methods. By comparing old CNN-based methods with new Transformer-based methods, this study tried to help researchers understand how object detection methods have changed over time. Thirteen important methods were looked at, which gave people faith in the growth of Transformers. At the end of the study, problems, chances, and possible futures in the field were summed up.

Francies et al.(2022) Three new YOLO algorithms (YOLOv3, YOLOv4, and YOLOv5) were used in a study to look at how well they could recognize 3D objects on a large dataset. With a mAP of 77% and an IOU of 0.41, YOLOv3 was the most accurate, but it ran for almost 8 hours longer. YOLOv4 had an IOU of 0.035, a mAP of 55%, and could run for 7 hours. Processing was faster for YOLOv5, which had a mAP of 48% and an IOU of 0.045. It took about 3 hours. A changed version of YOLOv5 that was tuned for hyperparameters and layers got a 55% mAP score and ran for 3 hours, showing that it was more efficient than other versions. The study found that YOLOv3 had the best recognition accuracy and that the suggested modified YOLOv5 had the fastest processing time.

Arulprakash et al.(2022) This highlighted the significance of object detection in computer vision, exploring deep learning concepts, CNN architectures, and evaluation methods. It discussed one-stage and two-stage recognition frameworks, considering factors like multi-scale variations and security. The conclusion provided essential steps for building effective object detectors and suggested future research directions.

Van dyck et al.(2021) For recognizing objects, Deep Convolutional Neural Networks (DCNNs) and the ventral visual pathway were put up against each other. Eye tracking was used to show changes in visualization methods in a study with 45 human observers and three DCNNs. The DCNN vNet, designed with biologically plausible receptive fields, closely matched human viewing behavior compared to a standard ResNet. There was a link between agreement on spatial object recognition and picture-specific factors like category, animacy, arousal, and valence, but not with difficulty or general image properties. At the point where biological and computer vision research meet, this work gives us new information.

Aziz et al.(2020) This study gave a thorough look at the latest progress made in using deep learning to find objects in pictures. It was divided into three main groups: methods based on area proposals, methods for recognizing and classifying objects, and new detectors. The main areas of focus were surveillance, defense, transportation, medical, and everyday uses. The poll looked at things that affect how well detection works and came up with fifteen current trends and ideas for where future research should go.

Liu et al.(2020) This paper gives a short summary of recent progress made in using deep learning to find objects. The collection of more than 300 research papers looks at many topics, including frameworks for detection, representing object features, making proposals, context modeling, training strategies, and evaluation measures. The study shows how deep learning has changed the way computer vision finds objects and offers possible directions for future research.

Lee et al.(2020) Object detection was a very important part of computer vision. Since 2012, there has been a lot of study using convolutional neural networks and modified structures. Object recognition problems, especially CNN's bounding box problems, were solved in large part by representative algorithms like convolutional neural networks and YOLO. The study showed two sets of algorithms based on CNN and YOLO and compared how well they worked in terms of speed, accuracy, and cost. In hindsight, YOLO v3 was praised for striking a good balance between speed and accuracy when compared to the most recent advanced answer.

Cao et al.(2019) The paper presented enhancements to the Faster R-CNN object detection method, addressing challenges in recognizing small objects. These improvements included a refined loss function, enhanced regions of interest pooling, and multi-scale convolution feature fusion. The proposed algorithm demonstrated superior performance, achieving a 90% memory rate and an 87% accuracy rate, particularly excelling in detecting small objects like traffic signs.

Zhao et al.(2019) Object detection research has evolved from traditional methods to advanced convolutional neural networks (CNNs) in recent years. CNNs excel in learning complex features, surpassing older approaches. The study reviews the history of deep learning and CNNs, explores diverse object detection architectures, and provides insights for improved performance. Task-specific needs are taken into account when covering things like detecting noticeable objects, faces, and pedestrians. Comparing methods is easier with the help of experimental studies. The paper ends with ideas for more study into object detection and related neural network-based learning systems.

Du et al.(2018) Since 2012, Convolutional Neural Networks (CNN) have made a lot of progress in object detection, which is an important part of picture processing. The transition to Faster R-CNN resulted in a mAP of 76.4, but its FPS remained slow (5 to 18), prompting a need for speed improvement. This paper explored You Only Look Once (YOLO), a CNN variant that broke from tradition, introducing a simpler and highly efficient approach to object detection. YOLO beat Faster R-CNN with an amazing FPS of 155 and a mAP of 78.6. Compared to the latest solutions, YOLOv2 struck an excellent balance between speed and accuracy, showcasing strong generalization for whole-image representation.

Tobías et al.(2016) showed how important Deep Learning (DL), especially Convolutional Neural Networks (CNN), is for jobs that need to recognize patterns. Using powerful General Purpose Graphic Processor Units (GPGPU) sped up the process of fixing problems, making it possible to build bigger networks with less computer time. Modern CNNs were able to do jobs like character recognition, face recognition, and object detection as well as humans. Advancements also empowered mobile devices to execute CNN models in real-time. The study focused on implementing lightweight CNN schemes for domain-specific object recognition on mobile devices.

Table : Comparison Table For The Author's Review .

Authors and Year	Main Focus	Key Results
Archana et al. (2024)	Image Processing, DL Models	PSPNet, Mask-RCNN for accurate object segmentation; Emphasized need for dependable image processing.
Ouf et al. (2023)	Smart Farming, YOLOv4	YOLOv4 - 98.52% mAP, fast recognition in leguminous seed detection.
Arkin et al. (2023)	Object Detection Evolution, CNN vs. Transformer	Transformer-based methods show potential in object detection tasks.
Francies et al. (2022)	3D Object Recognition, YOLO Algorithms	YOLOv3 - best recognition accuracy; Modified YOLOv5 - efficient processing time.
Arulprakash et al.(2022)	Object detection in computer vision, deep learning concepts, CNN architectures, and evaluation methods	Explored one/two-stage recognition frameworks, considered multi-scale variations and security. Provided steps for effective detectors, suggested future research.
Van dyck et al. (2021)	DCNN vs. Ventral Visual Pathway	vNet DCNN aligns with human viewing behavior; Image-specific factors linked to spatial object recognition.
Aziz et al.(2020)	Recent deep learning advancements in object detection, covering region proposal-based and recognition methods	Addressed applications in surveillance, military, transportation, medical, daily life. Discussed detection factors, trends, future research.
Liu et al.(2020)	Recent advancements in	Encompassed 300 papers,

	object detection using deep learning methods, exploring various aspects	highlighted transformative impact. Suggested potential directions for future research.
Lee et al. (2020)	Object Detection, CNN, YOLO	YOLO v3 - favorable speed-accuracy trade-off in object detection.
Cao et al.(2019)	Enhanced Faster R-CNN for small object detection	Proposed refined loss function, enhanced pooling, multi-scale fusion. Achieved 90% memory, 87% accuracy, excelling in small object detection like traffic signs.
Zhao et al.(2019)	Evolution of object detection research, history of deep learning, and CNNs	Explored diverse architectures, provided insights for improved performance. Covered tasks like salient object, face, and pedestrian detection. Suggested future research directions.
Du et al. (2018)	Object Detection Progress, YOLO	YOLOv2 - High FPS (155), mAP (78.6), balanced speed and accuracy.
Tobías et al. (2016)	DL, CNNs in Pattern Recognition	CNNs achieved human-comparable performance; Mobile devices enabled real-time execution.

4. Datasets And Evaluation Metrics

The prominent datasets that are most frequently used for object detection tasks are described in this section Fig 2. The dataset for object detection is download from cocodataset.org site. The dataset consists of images of different 80 classes. The Microsoft Common Objects in Context (MS-COCO) collection [20] is one of the hardest ones out there. Ever since its launch in 2015, its popularity has only grown. There are over two million occurrences, with each image including an average of 3.5 categories. Additionally, it is the most popular dataset compared to others, with 7.7 instances per image. MS COCO also includes photos taken from a variety of angles. The dataset consists of many objects' classes.

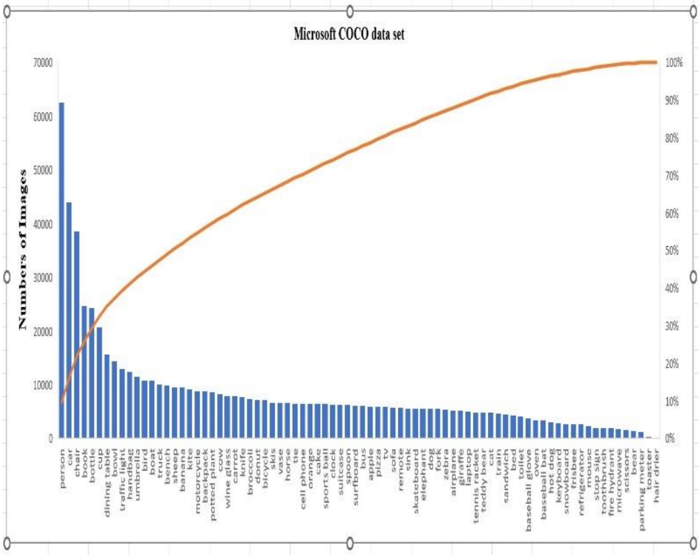


Fig.2.Number of different classes annotated in MS COCO dataset.

4.1 Metrics

Object detectors measure their efficiency in three ways: frames per second (FPS), precision, and recall. Mean Average Precision, or mAP[21], is the most common tool used for review, though. The source of accuracy is Intersection over Union (IoU), which is the ratio of the area of union to the area of overlap between the real-world box and the forecast box. To ascertain whether the detection is accurate, a threshold is chosen. An IoU below the cutoff is called a False Positive, and one above it is called a True Positive[22]. When a model can't find an object that is in the ground truth, this is called a false negative. Recall assesses accurate predictions in relation to ground truth, whereas precision indicates the fraction of correct forecasts.

- Each class's average precision is calculated independently using the algorithm above. Mean average precision (mAP), which is the average of all the classes' average precisions, is used as the single measure for the final test to compare how well the detectors worked.
- **Accuracy**

  It serves as the primary criterion for assessing a model's efficiency. An improved model will outperform the other in terms of accuracy. An equation is used in the calculation.

$$\mathbf{Accuracy} = \frac{\mathbf{TP + TN}}{\mathbf{TP + TN + FP + FN}} \qquad \mathbf{(1)}$$

falsely reported positive cases (FP), falsely reported negative cases (FN), true positive values (TP), true negative values (TN), and falsely reported positive cases (FP).

- **F1-score**

  It is another popular way to measure success and is found by taking the harmonic mean of recall and precision. To determine the F1-score, an equation is used.

$$\mathbf{F1} = \frac{\mathbf{2 \times Precision \times Recall}}{\mathbf{Precision + Recall}} \qquad \mathbf{(2)}$$

where these formulas are used to compute Precision and Recall.

$$\mathbf{Precision} = \frac{\mathbf{TP}}{\mathbf{TP + FP}}$$
$$\mathbf{Recall} = \frac{\mathbf{TP}}{\mathbf{TP + FN}} \qquad \mathbf{(3)}$$

True positive (TP) means "yes"; true negative (TN) means "no"; false negative (FN) means "yes"; and so on.

5. Backbone Architecture

The EfficientNet The simple but effective way that Dang et al. [23] use to scale up models is called the compound coefficient. Instead of randomly growing width, depth, or resolution, compound scaling always uses the same set of scaling coefficients on all dimensions. While scaling a single dimension can help improve model performance, EfficientNet found that scaling the width, depth, and image resolution while taking into account the changeable available resources balances the scaling in all three dimensions and increases overall model performance. In [24], Tan et al. did a full study on network scalability and how it affects model performance. They gave an overview of how different network factors, like depth, width, and resolution, affect how accurate the results are. In their example, they showed that scaling each number separately has costs. Moving deeper into a network can help it understand more complicated and rich features, but it can be hard to train these networks because of the vanishing gradient problem. In a similar vein, increasing network width will facilitate the acquisition of finer information at the expense of posing challenges in getting higher level data. When you increase the picture resolution, things like width and depth become more useful as the model gets bigger. In the document [25], Tan et al. recommended using a compound coefficient that could scale all three dimensions in the same way.

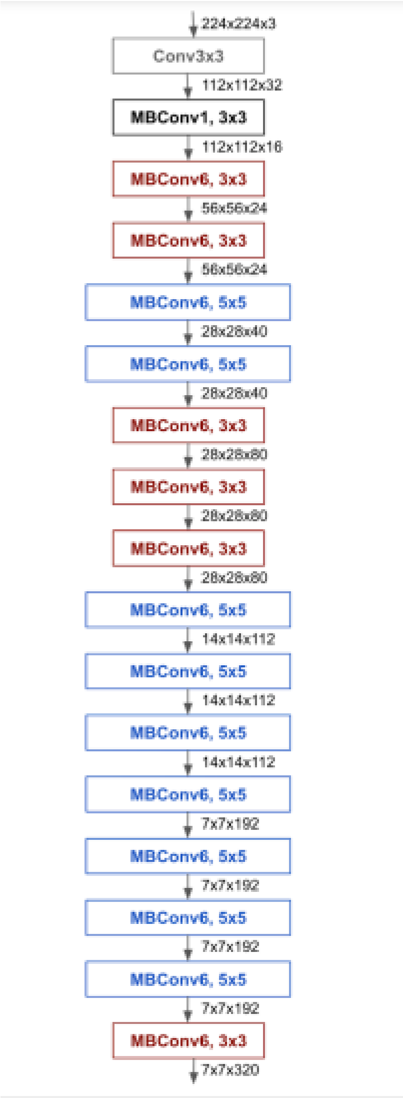


Fig. 3 Architecture of EfficientNet (Dang Thi Phuong Chang 2017)

It is possible to find the constant that goes with every modal value by grid searching a baseline network with the coefficient set to 1. The foundational architecture, influenced by their earlier efforts [26], is created by neural architecture search that maximizes computations and accuracy on a search target. The design of EfficientNet is straightforward and effective. The old model worked slower and less correctly than the new one, but it was much smaller. It could be the start of a new era in the study of networks that work well because it makes networks work much better. Figure 3 provides a visual illustration of EfficientNet.

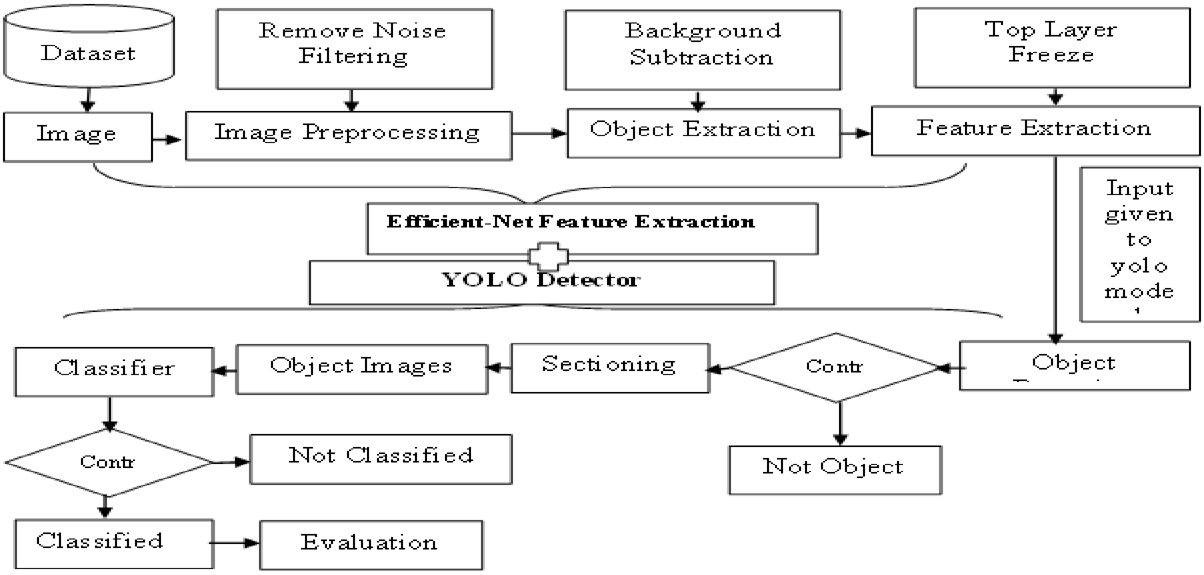


Fig. 4Flow Diagram of Hybrid Approach (E-YOLO)

6. Object Detector

Object detection as a classification problem is solved by two stage detectors; a module offers candidates that the network identifies as background or objects. But You Only Look Once, or YOLO [27], reframed it as a regression problem, figuring out directly what the picture pixels are and how big their bounding boxes are. In

YOLO, the image you send is split into a S×S grid, and the item is found in the cell where its center falls. A grid cell predicts multiple bounding boxes, and each prediction array consists of 5 elements: center of bounding box – x and y, dimensions of the box – w and h, and the confidence score. YOLO was much more accurate and faster than single stage real time models that were available at the time..

7. Proposed Model

The proposed system Fig 4 employs EfficientNet + YOLO hybrid technique to recognise and detect the objects. The main goal of the suggested system is to take an image as input and put it into a category with a labeled class name that shows where it belongs. CNN is mostly used to pull out features and label names based on classification. Here is the Architecture design of the proposed model.

1. Trained EfficientNet model and freeze its layer to void any weight updates.
2. Remove the top layer of EfficientNet model to obtain the extracted Features.
3. Create YOLO model adds the extracted feature as input to the YOLO model.
4. Train the Hybrid model to the dataset.
5. Evaluate the model’s performance using standard evaluation metrics.

Through Hybrid approach experiments, the following improvements are:

- The Efficient-Net [28] series B0 backbone feature extraction network reduces the size of the network model's parameters without affecting the accuracy of the model.
- On the other hand, using the 9x9 and 13x13 spatial pyramid pooling structure makes the model's receptive field and detecting accuracy better.

- The bidirectional feature pyramid structure is used instead of the original algorithm's unidirectional feature pyramid to improve feature extraction from the network and improve the meaningful information of the output feature layers at different sizes.
- COCO datasets, which contain 80 labelled classes, were used in the development of object identification and classification systems. We employed EfficientNet Neural Network Models for categorization. It was possible to find multiple objects in the picture by using YOLO deep learning tools.
- In order to enable the proposed deep learning-based object recognition model to handle real-world issues across a variety of domains, hybrid object recognition E-YOLO (B0EfficientNet+ YOLOv8) algorithms that can be created for the recognition of an object in the image are applied. By using this technique, the suggested model was trained and tested on diverse sets of photos featuring distinct items.
- • The image that YOLO Figure 5 is given is split into S × S grids, and the item is found in the cell that has its center in the middle. A grid cell is sure to guess several boundary boxes. At finally, the class probability map predicts the item.

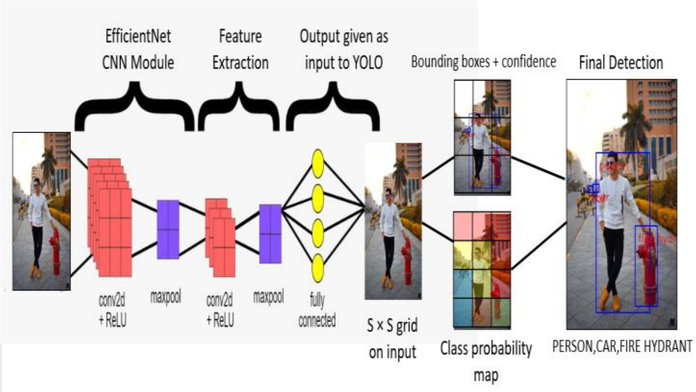


Fig.5. Illustration of the internal architecture of proposed model

8. Experiment And Discussion

A bunch of tests were done to see how well the suggested system worked and how well it compared to other object detection methods that are already out there.

COCO dataset was used in experiments. This is public dataset collected by Microsoft COCO a few year ago.it contain thousands of color images each annotated with 80 objects categories. Every picture is in the JPG format and has a set resolution of 640 x 480. Figure 2 displays the distribution of the photos among the 80 classes. RGB version of the picture, where each colour channels contain 8 bits per pixel.

Each experiment divided the dataset into two sections: training and testing, following preprocessing. Eighty percent of the data set stated above is used to train the model, and twenty percent is used to test it. 5% of the training data set was also used for confirmation during the fitting process.

Keras's SGD (stochastics gradient descent) was the optimizer we chose because it has an adjustable learning rate (±). Each epoch has a different learning rate number. The values of it depend on the epoch numbers as

$$\boldsymbol{\alpha_n = \alpha_0 \times 0.1^{ep(n)}/_{10}} \qquad \textbf{(4)}$$

What we have here is ep(n), α_0 is the learning rate at the beginning, and α_n is the learning rate at epoch number n.

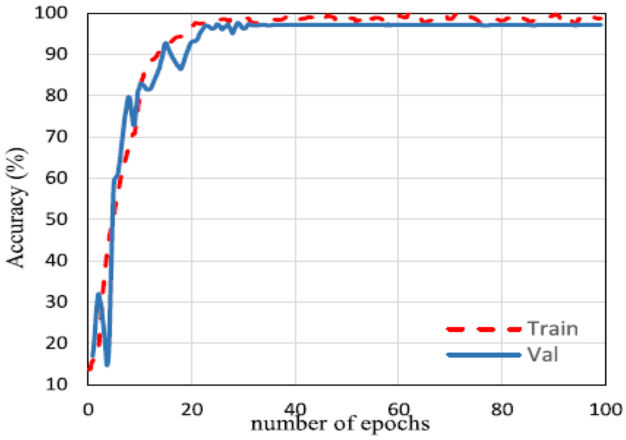


Fig.6. Training and validation accuracy curve

There were 80 classes and 100 iterations used to fit the model to the training sample. The suggested model's success was tested in our experiment. On the test data, we obtained a 97.82% accuracy rate. Figure 6 shows how the training and confirming accuracy change with the number of epochs. The training and validation accuracy curves demonstrate a distinctly noticeable improvement in the model's accuracy.

The platform on which the experiment is conducted is as follows: Operating System: Windows 11 Home; Processor: Intel® Core (TM) i5-1135G7 11th generation @ 2.4GHZ; Memory: 8GB; GPU: Intel® Iris ® Xe Graphics; Program: karas environment using Python Language implementation on Jupyter Notebook.

9. Comparative Result

Utilizing Microsoft COCO datasets, we test how well various object analyzers work. Object detector performance is affected by many things, such as the input picture size and scale, the feature extractor, the GPU architecture, the number of proposals, the training method, the loss function, and more. This makes it hard to compare different models without a common benchmark setting. The average precision (AP), recall, F1-score, and accuracy of the models are used to compare them at the inference time. When the IoU of the projected bounding box with the ground truth is higher than 0.5, the average accuracy for all classes is reached, which is called AP (0.5). When feasible, we purposefully compare detector performances on input images of comparable sizes in order to offer a plausible explanation.

The proposed work employs E-YOLO classifier for precision in classification and recognition for objects. The percentages of recall, accuracy, precision, and f1-score from the proposed task are shown in Table 2-4. Together with a comparative of others classifier VGG-16, ResNet101, Hourglass, Resnet101Detr.

Table.2. Precision

Model	**Backbone**	**A.P[0.5] (%)**
SSD	VGG-16	41.2
RetinaNet	ResNet-101-FPN	49.5
CenterNet	Hourglass-104	61.1
Detr	ResNet-101	63.8
YOLO	**EfficientNet-B0**	**90.75**

Table.3. Recall

Model	**Backbone**	**Recall (%)**
SSD	VGG-16	52
RetinaNet	ResNet-101-FPN	60
CenterNet	Hourglass-104	75.76
Detr	ResNet-101	83.21
YOLO	**EfficientNet-B0**	**95.26**

Table.3. F1-Score

Model	**Backbone**	**F1-Score (%)**
SSD	VGG-16	46
RetinaNet	ResNet-101-FPN	54.2
CenterNet	Hourglass-104	67.64
Detr	ResNet-101	72.22
YOLO	**EfficientNet-B0**	**92.95**

Table.4. Accuracy

Model	**Backbone**	**Accuracy (%)**
SSD	VGG-16	72.6
RetinaNet	ResNet-101-FPN	80.22
CenterNet	Hourglass-104	89.5
Detr	ResNet-101	92.13
YOLO	**EfficientNet-B0**	**97.82**

From graphical Representation Fig 7 it is clear that E-YOLO shows the highest bar in terms of all scales that is in average precision its value is 90.75, in recall it shows value 95.26, in F1- score 92.95 and the proposed model achieved an object recognition accuracy of 97.82% which is highest of all values in comparison with other models.

List of detected objects with their bounding boxes and class labels as shown in Table 5 byhybrid classifier E-YOLO.

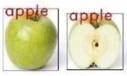
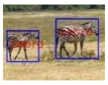
				
Airplane	Apple	Tennis racket	Potted plant	Banana
				
Baseball bat	Bear	Horse	Zebra	Bed
Person, Bench	Bicycle	Person	Boat	Bird
Wine glass	Mouse	Chair	Clock	Kite, person

Table.5. Detected Objects

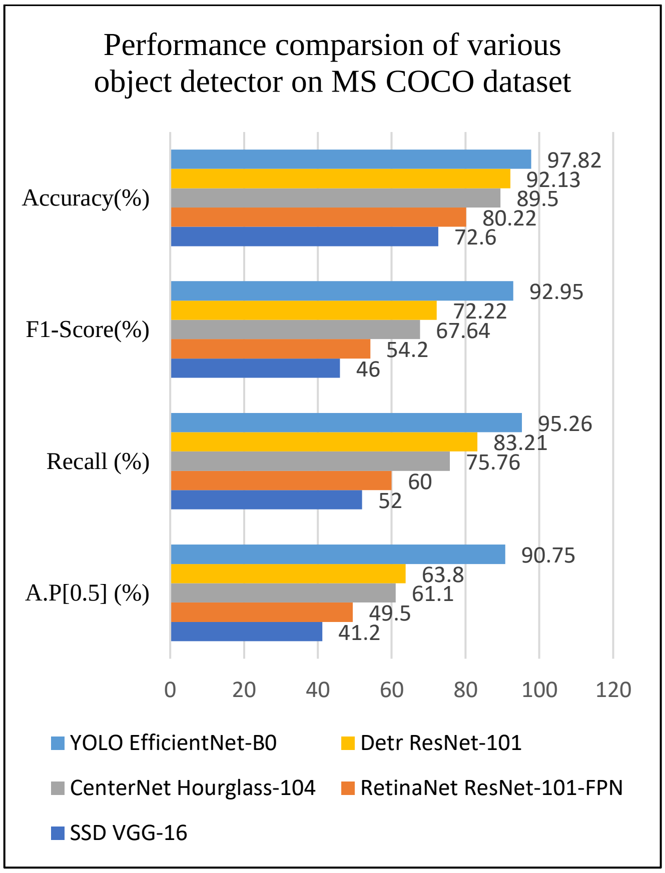


Fig.7. Performance comparison of various object detector on MS COCO dataset

10. Conclusion

This research suggested a deep learning-based object recognition framework. EfficientNet CNN model play important role as a backbone to trained the proposed model with the using of YOLO detector for the detecting objects. For the training and testing purpose MS COCO dataset is used. This hybrid approach named E-YOLO hasachieved excellent accuracy on dataset. The suggested model's performance has been contrasted with existing model and gives 97.82 % accuracy which is much better than others. To keep the same level of speed while adding more classes, researchers plan to do more work in the near future.

References

[1] Van De Sande, K., Gevers, T., & Snoek, C. (2009). Evaluating color descriptors for object and scene recognition. IEEE transactions on pattern analysis and machine intelligence, 32(9), 1582-1596.

[2] Kuang H, Hang Chan LL, Liu C, Yan H, *"Object detection and object classification based on weighted scorelevel feature fusion",* J Electron Imaging ,2016; 25:1–11.

[3] Hossain MS, AlHammadi M, Muhammad G, "*Automatic object detection and object classification using deep learning for industrial applications"*, IEEE Trans Industry Inf,2019 15(2):1027–1034. 2019.

[4] Jiang L, Koch A, Scherer SA, Zell A, *"Multiclass object detection and object classification using RGB-D data for indoor robots"*, IEEE Int Conf Robot Biomimetics, Shenzhen, China, 2013.

[5] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, M. Pietikäinen, *Deep learning for generic object detection: a survey*, Version: 1, arXiv:1809 .02165, 2018.

[6] Yu, J., Guo, K., Hu, Y., Ning, X., Qiu, J., Mao, H., ... & Yang, H. (2018, March). Real-time object detection towards high power efficiency. In 2018 Design, Automation & Test in Europe Conference & Exhibition (DATE) (pp. 704-708). IEEE.

[7] Archana, R., & Jeevaraj, P. E. (2024). Deep learning models for digital image processing: a review. Artificial Intelligence Review, 57(1), 11.

[8] Ouf, N. S. (2023). Leguminous seeds detection based on convolutional neural networks: Comparison of faster R-CNN and YOLOv4 on a small custom dataset. Artificial Intelligence in Agriculture, 8, 30-45.

[9] Arkin, E., Yadikar, N., Xu, X., Aysa, A., & Ubul, K. (2023). A survey: object detection methods from CNN to transformer. Multimedia Tools and Applications, 82(14), 21353-21383.

[10] Francies, M. L., Ata, M. M., & Mohamed, M. A. (2022). A robust multiclass 3D object recognition based on modern YOLO deep learning algorithms. Concurrency and Computation: Practice and Experience, 34(1), e6517.

[11] Arulprakash, E., & Aruldoss, M. (2022). A study on generic object detection with emphasis on future research directions. Journal of King Saud University-Computer and Information Sciences, 34(9), 7347-7365.

[12] Van Dyck, L. E., Kwitt, R., Denzler, S. J., & Gruber, W. R. (2021). Comparing object recognition in humans and deep convolutional neural networks—an eye tracking study. Frontiers in Neuroscience, 15, 750639.

[13] Aziz, L., Salam, M. S. B. H., Sheikh, U. U., & Ayub, S. (2020). Exploring deep learning-based architecture, strategies, applications and current trends in generic object detection: A comprehensive review. IEEE Access, 8, 170461-170495.

[14] Liu, C., Tao, Y., Liang, J., Li, K., & Chen, Y. (2018, December). Object detection based on YOLO network. In 2018 IEEE 4th information technology and mechatronics engineering conference (ITOEC) (pp. 799-803). IEEE.

[15] Lee, Y. H., & Kim, Y. (2020). Comparison of CNN and YOLO for Object Detection. Journal of the semiconductor & display technology, 19(1), 85-92.

[16] Cao, C., Wang, B., Zhang, W., Zeng, X., Yan, X., Feng, Z., ... & Wu, Z. (2019). An improved faster R-CNN for small object detection. Ieee Access, 7, 106838-106846.

[17] Zhao, Z. Q., Zheng, P., Xu, S. T., & Wu, X. (2019). Object detection with deep learning: A review. IEEE transactions on neural networks and learning systems, 30(11), 3212-3232.

[18] Du, J. (2018, April). Understanding of object detection based on CNN family and YOLO. In Journal of Physics: Conference Series (Vol. 1004, p. 012029). IOP Publishing.

[19] Tobías, L., Ducournau, A., Rousseau, F., Mercier, G., & Fablet, R. (2016, December). Convolutional Neural Networks for object recognition on mobile devices: A case study. In 2016 23rd International Conference on Pattern Recognition (ICPR) (pp. 3530-3535). IEEE.

[20] Raparthi, M., Dodda, S. B., & Maruthi, S. (2023). Predictive Maintenance in IoT Devices using Time Series Analysis and Deep Learning. Dandao Xuebao/Journal of Ballistics, 35(3). https://doi.org/10.52783/dxjb.v35.113

[21] A. Geiger, P. Lenz, C. Stiller, R. Urtasun, *Vision meets robotics: the KITTI dataset*, Int. J. Robot. Res. (2013).

[22] Pal, S. K., Pramanik, A., Maiti, J., & Mitra, P. (2021). Deep learning in multi-object detection and tracking: state of the art. *Applied Intelligence*, *51*, 6400-6429.

[23] Cai, Z., & Vasconcelos, N. (2018). Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6154-6162).

[24] Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., & Pietikäinen, M. (2020). Deep learning for generic object detection: A survey. International journal of computer vision, 128, 261-318.

[25] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, L. Fei-Fei, ImageNet large scale visual recognition challenge, Int. J. Comput. Vis. 115(3) (2015) 211–252, https://doi .org /10 .1007 /s11263 -015 -0816.

[26] M. Tan, Q.V. Le, EfficientNet: rethinking model scaling for convolutional neural networks, arXiv:1905 .11946.

[27] M. Tan, B. Chen, R. Pang, V. Vasudevan, M. Sandler, A. Howard, Q.V. Le, Mnas-Net: platform-aware neural architecture search for mobile, https://arxiv.org /abs /1807.11626v3.

[28] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp.779–788.

[29] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, M. Pietikäinen, *Deep learning for generic object detection: a survey*, Int. J. Comput. Vis. 128(2) (2020) 261–318.