DOI: https://doi.org/10.33103/uot.ijccce.23.2.11

Object Detection Using Deep Learning Methods: A Review

Asmaa Hasan Alrubaie[1], Maisa'a Abid Ali Khodher[2], Ahmed Talib Abdulameer[3]
[1]*Computer Science Department, University of Technology, Baghdad, Iraq*
[2]*Computer Engineering Department, University of Technology, Baghdad, Iraq*
[3]*IT Department, Technical College of Management, Middle Technical University, Baghdad, Iraq*
[1]*cs.20.41@grad.uotechnology.edu.iq*, [2]*Maisaa.A.Khodher@uotechnology.edu.iq*,
[3]*ahmed.talib@mtu.edu.iq*

Abstract— *Target detection, one of the key functions of computer vision, has grown in importance as a study area over the past two decades and is currently often employed. In a certain video, it seeks to rapidly and precisely detect and locate a huge amount of the objects according to redetermined categories. The two forms of deep learning (DL) algorithms that are used in the model training algorithm are single-stage and 2-stage algorithms of detection. The representative algorithms for every level have been thoroughly discussed in this work. The analysis and comparison of numerous representative algorithms in this subject is after that explained. Last but not least, potential obstacles to target detection are anticipated.*

Index Terms— *Object detection, Deep learning, Regions of interest (ROI), Convolutional Neural Networks (CNNs).*

I. INTRODUCTION

Numerous aspects of daily life already include artificial intelligence (AI), including software for predictive analysis, self-driving cars, and face recognition. Object detection is a key area in AI that draws a lot of interest. One-and two-stage object detection algorithms, the two depend on DL techniques, make up the majority of the present popular object detection algorithms. The key difference between these two approaches is whether or not a region proposal is generated. The generation of a region proposal is not necessary for one-stage object detection algorithms. It has immediate access to the object's coordinate position and classification accuracy. But two-stage object detection algorithms must produce a region proposal to classifying and locating the object[1].

Window sliding, preprocessing, feature selection, post processing, feature extraction, and feature classification represent the 6 basic processes in conventional detection algorithms for manually extracting features, which are typically used for specialized recognition tasks. Low portability, small data size, high time complexity, window redundancy, lack of pertinence, lack of robustness for changes in variety, and acceptable performance only in certain simple cases are the key drawbacks [2].

Because multi-objects might appear in videos in various poses and with varied characteristics, it could be difficult to detect them. Because of this, it is crucial to build a reliable system for motion detection and recognition. In the present study, the objective is to locate moving objects using a convolutional network. To effectively recognize objects, it

Received 25/September/2022; Accepted 08/December /2022
 ISSN (Print) 1811-9212 *ISSN (Online) 2617-3352*

DOI: https://doi.org/10.33103/uot.ijccce.23.2.11

employed a variety of DL algorithms. The outcome of this step is object localization, which entails drawing a bounding box around at least one object in a video[2].

The following is how the paper will be organized: Section II deals with the related works in the object detection using traditional and deep learning methods, Section III explains the performance comparison of various algorithms, and finally Section IV presents the conclusion of this paper.

II. RELATED WORK

The current literature concerning the object detection using traditional and deep learning methods suggests a shortage of active research for this area; there are only a couple of publications that combined traditional and deep learning methods to object detection. Hence, it broadened our literature review to focus on the object detection using traditional and deep learning methods. The literature on object detection is vast, and in this section we will focus on approaches exploiting class-agnostic ideas and addressing scalability.

In [2019], Christian Szegedy, et al, presented DL algorithm to find moving objects. Deep Neural Networks (DNNs) have lately demonstrated remarkable performance on object detection with the use of DNNs, through precisely localizing objects. This work addresses the challenge of detection and positioning of moving objects. It describes an approach to object detection as regression problem to object bounding box that is both straightforward and effective. When it comes to object detection, DNNs show significant differences from conventional methods. Since they are deep architectures, they can learn more complex models compared to shallow ones [3].

In [2020], Juncai Zhu, et al, proposed the motion information in the image is identified using the background compensation technique. The binary mask regarding moving regions has been acquired via employing inter-frame difference approach after the motion parameter has been determined based on coordinate relation of the feature points in the neighboring frames. You Only Look Once v3 - Segregation of Duties (YOLOv3-SOD) network is used in the DL approach, which tries to find moving targets more adequately. When put to comparison with previous techniques, this approach performed better and was able to detect moving targets with greater accuracy. This approach still has certain drawbacks, though. Particularly, the motion detection module's performance is subpar [4].

In [2021], Sankar K. Pal, et al, suggested the object detection and tracking according to DL framework, the problem of labelling various objects in the image frame with their precise classes and accurately anticipating their bounding boxes can be seen as the object recognition problem. DL is thus computationally demanding and challenging to engineer. In order to enable extremely quick motion detection and object recognition, a high-performance graphics processing units (GPU) is needed. One-stage detectors are capable of filtering out the simple samples by properly setting the focal loss function. This considerably reduces the number of the target proposals and boosts speed and accuracy of detection. The same might be true for two-stage detectors. In comparison to find product by individuals, combining two-stage and one-stage detectors yields better outcomes [5].

In [2022], Mallineni Priyanka, et al, , presented the object detection and classification became achievable with the emergence of new rising DL technologies. In feature to

Received 25/September/2022; Accepted 08/December /2022
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.23.2.11

conventional object detection techniques, DL techniques are capable of learning and rendering features. Faster R-CNN is chosen as the best option since it is more precise and economical compared to the R-CNN, yet CNN models could just be utilized for image classification and cannot localize objects. A Faster R-CNN is employed for the task of object detection, which presents to users as one, unified network from start to finish. It could as well specify positions regarding different things accurately. Those models have been considered as the most accurate, although they're often slow [6].

III. PERFORMANCE COMPARISON OF VARIOUS ALGORITHMS

Currently, object detection is a rather well-liked field. From the conventional techniques to the DL techniques. In this study, object detection algorithms are reviewed. The working principles of each algorithm are explained in depth, and the differences and similarities between them are examined. The efficiency of each algorithm is put to comparison with the experimental data, as listed in Table I.

TABLE I. OBJECT DETECTION ALGORITHMS

Methods	year	Architecture	Advantage	Disadvantage	Method Used
Traditional	2014	The fundamental design of conventional algorithms of object detection. The region selector primarily makes use of sliding windows with various ratios and sizes to slide on image from the left to the right and top to bottom by a specific step size. The feature extractor primarily uses the Haar, HOG, SIFT, and other algorithms. Lastly, the classifier determines object category algorithms like Adaboot and support vector machine (SVM).	DL is occasionally overkill because conventional Computer Vision approaches may frequently solve a problem considerably more quickly and with less code compared to DL. The performance of algorithms like SIFT, as well as very basic color thresholding and pixel counting algorithms, is not class-specific, making them incredibly universal and applicable to any image.	- The problem with the conventional method is that it requires expensive processing resources to generate candidate bounding boxes with the use of sliding window method. - Not all types of objects can be perfectly described by their engineered features.	1) Deformable Part Models (DPM) 2) Histogram of Oriented Gradients (HOG) Detector 3) Viola-Jones Detector
RCNN	2014	The architecture of RCNN is divided into three stages: 1) Regional Proposal Generation. 2) Feature Extraction. 3) Localization and classification.	It makes candidate bounding boxes of higher quality and extracts high-level features using deep architecture.	RCNN modest training set will take a long time to process with very deep networks like Visual Geometry Group VGG16.	Two-stage detection
SPPNet	2014	SPPNet's strength is crucial for object detection. It divided image into (sub-images) for the training of detectors through	SPPNethas improved outcomes by accurately estimating various region proposals at their appropriate scales,	The drawback of SPPNet is the same as that of RCNN, therefore additional storage space costs and time are still	Two-stage detection

Received 25/September/2022; Accepted 08/December /2022
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.23.2.11

		initially calculating feature maps from compute image once utilizing SPPNet. Through using this technique, convolutional features are not computed repeatedly.	and it also increases detection effectiveness during testing periods by distributing the cost of calculation prior to the SPP layer among various proposals.	necessary.	
Fast RCNN	2015	Fast R-CNN uses a conventional convolution architecture like VGG16 to process the full image for the purpose of producing a feature map similar to SPP-Net.	The training for all of the network layers in the Fast R-CNN might be completed in one step with a multi-task loss. It lowers the price of additional storage space and improves precision and effectiveness.	Similar to RCNN, Fast-RCNN uses selective search to find the ideas for the region. The selective search slows down the performance of the network and is a time-consuming process.	Two-stage detection
Faster RCNN	2015	Instead of using selective search approach on feature map, a separate network has been used in order to anticipate the region proposals in the faster RCNN.	With an aid of Faster R-CNN proposal, region proposal-based CNN models for object detection might truly be trained from beginning to end.	The training process takes a long time.	Two-stage detection
YOLO	2015	YOLO algorithm utilizes the following 3 approach: 1) Residual blocks: initially, the image is separated to several grids. 2) Regression of the bounding box: An outline drawing attention towards an object in an image has been referred to as a bounding box. 3) Intersection over union (IOU): which represents an object detection phenomenon which explains the way that the boxes overlap.	The next reasons are why YOLO is so attractive: 1) Speed: Since it could anticipate objects in real time, the speed of the detection has increased. 2) High accuracy: the outcomes with a small number of the background errors. 3) Learning capabilities: due to the outstanding learning abilities of YOLO, it has the ability to learn object representations and use them in object detection.	1) Compared to the R-CNN family of algorithms, object detection is faster because it only requires one step, although it occasionally displays worse accuracy. 2) YOLO finds it challenging to handle small objects in groups.	One-stage detection
SSD	2016	An SSD head and a backbone model make up SSD architecture. A pretrained network of image classification	1) the detect objects at various scales and produce a tighter bounding box thanks to the	Single Shot Detector (SSD) is quick, yet accuracy suffers as a result.	One-stage detection

Received 25/September/2022; Accepted 08/December /2022

 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.23.2.11

		plays the role of a backbone model's feature extractor in the majority of the cases. The SSD head is only one or more convolution are added to that backbone, with outputs being read as bounding boxes.	SSD design. 2) Not all things have a square shape. To varied degrees, a few are wider and a few are longer. The SSD architecture enables the anchor boxes'.		
Mask R-CNN	2017	Masking the R-CNN model has been divided to 2 parts, which are: 1) RPN to suggested candidate object bounding boxes. 2) Binary mask classifier for the generation of the mask for each one of the classes	Anchor boxes are used by Mask R-CNN for the detection of multiple objects, objects of various sizes, and objects that overlap in an image. Which increases effectiveness and speed of object detection.	The region proposal-based frameworks are time-consuming and might not be appropriate for real-time applications, as it witnessed with Mask R-CNN.	Two-stage detection
Retina Net	2017	An architecture for a Retina Net model consists of four main parts: a) Bottom-up Pathway. b) Top-down path-way and Lateral connections. c) Classification subnetwork. d) Regression subnetwork.	A very good one-stage object detection model, Retina Net had shown to be effective with the dense and small objects. Which is why, it became popular as one of the object detection models to be used with the aerial and satellite imagery.	Larger backbone networks used by Retina Net produce higher accuracy, yet slower inference speeds. The training time lasts from 10 and 35 hours.	One-stage detection
YOLOv3	2018	YOLOv3's architecture separates an image first into a grid. The number of boundary boxes around the objects scoring highly in the abovementioned specified classes is predicted for every grid cell. Assuming that the prediction must be accurate, each boundary box has a corresponding confidence score and only detects one object per bounding box.	1) The bounding boxes and the class likelihoods for those boxes have been predicted by YOLO3 using a single convolutional network. 2) YOLO3 has the benefit of being far faster compared to other networks while keeping accuracy.	YOLO-v3 has been considered as a great substitute for the models if could be trained with large data-sets because it might be utilized for the detection of the objects with accuracy similar to that of Retina Net in the case of utilizing a larger data-set.	One-stage detection
YOLOv4	2020	The architecture is made up of several components. The input, which comes first, is essentially the set of	1) Due to how inaccurate YOLO has been at the detection of small objects, the	Yolov4 does have its disadvantage in object detection. In the case when objects in the image exhibit	One-stage detection

Received 25/September/2022; Accepted 08/December /2022
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.23.2.11

		training images that will be supplied to the network, and it is processed in parallel by GPU in batches. Neck and the Backbone that perform feature.	precision for the small objects in YOLOv4 has been unmatched.	uncommon ratio characteristics, it does not generalize well.	
G-RCNN	2021	Two elements make up the G-RCNN architecture. Foreground RoIs are provided across the video frame by the first section, Object classification in every one of the ROIs is the subject of the second section, which is referred to as the classification network.	1) By combining the new idea of granulation in the deep CNN, G-RCNN can be described as an enhanced version of the common Faster RCNN and Fast RCNN for the extraction of the ROIs. 2) G-RCNN has enhanced the accuracy and speed of real-time object detection.	It is challenging to debug and improve G-RCNN network since it is a two-stage network. Time spent on training and testing lengthens in the meantime.	Two-stage detection
YOLO v5	2021	YOLO v5 architecture processes the whole image using one NN, after that separates it to sections and forecasts bounding boxes and probability values for every one of the components. The predicted likelihood is weighted while determining such bounding boxes. The approach "only looks once" at the image.	1) YOLOv4 is around 88% smaller (27 MB vs 244 MB). 2) Compared to YOLOv4, it is around 180% faster frames per second (140FPS vs 50FPS).	Because YOLOv5 is still being developed and it frequently receive updates from ultralytics, there is a major problem. Future revisions to various settings could occur.	One-stage detection
YOLO v7	2022	Generally, YOLOv7 have more precise object detection performance, a more robust loss function, and an improved label assignment and model training efficiency. In comparison to other DL models, YOLOv7 needs computational hardware that is many times less expensive. Without any pre-trained weights, it could be trained significantly more quickly on small datasets.	1)YOLOv7 algorithm is the most recent YOLO algorithms outperforms all earlier object detection algorithms and YOLO versions in terms of speed. 2) Compared to YOLOv4, YOLOv7 uses 36% less computation, decreases the number of parameters by 75%, and produces 1.5% greater AP (average precision).	As each grid can only detect one object, YOLOv7 struggles to detect and separate small objects in images where they occur in groups. As a result, YOLOv7 has trouble detecting and Localizing small objects that ordinarily from groups, like a line of ants.	One-stage detection

Received 25/September/2022; Accepted 08/December /2022
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.23.2.11

A. Before 2014 – traditional object detection period

The primary technique prior to DL is the conventional target detection algorithm. The information region selection, classifier design, and feature extraction steps make up the conventional target detection technique. In the information region selection aspect, a multi-scale sliding window has been frequently utilized for scanning the entire image in order to obtain a big number of the target candidate regions. This practice produces a huge number of the redundant candidate regions, increasing the amount of the calculations. The outcome of the feature extraction directly influences the impact of target detection in the feature extraction aspect. Conventional target detection algorithms collect the feature data from the image using Viola-Jones Detector, HOG, DPM, and other algorithms of feature extraction, as explained below:

1. Viola-Jones Detector (2001), the pioneering study which started the development of the conventional techniques of object detection.
2. HOG Detector (2006), a significant feature descriptor for object detection in image processing and computer vision.
3. DPM (2008) with the first introduction of the bounding box regression.

Most of them employ conventional ML classification techniques to choose the retrieved features when it comes to classifier design. *Fig. 1* depicts the conventional target detection. First, the algorithm receives the image of an object to be tested, and utilizing region selection, creates a huge number of candidate boxes. The features are after that extracted using the feature extraction algorithm. The target may be recognized by the classifier and appropriately classified using the retrieved feature image. There are two drawbacks to the conventional target detection algorithm. One is that the production of many candidate boxes as a number of information region selection necessitates the use of numerous computational resources. Second, the variety of objects, background image, lighting conditions and weak generalization capacity will conflict with the conventional feature extraction algorithm [5].

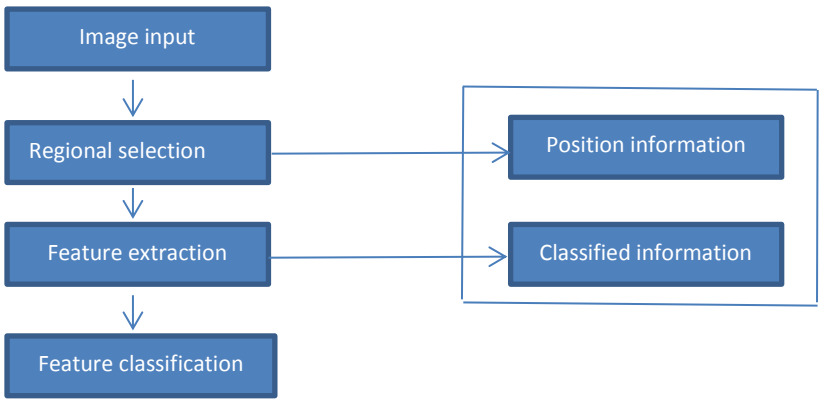


FIG. 1. TRADITIONAL FLOW OF THE TARGET DETECTION ALGORITHM.

B. After 2014 – deep Learning based object detection period

The advancement of DL is accelerated by increased computer performance, ushering in a new stage of AI. DL has been steadily used to image detection and recognition recently, which has tremendously aided in the advancement of this field's study. The performance of computer processing is expanding explosively as a result of the onset of the digital era, [7]. Deep CNNs have been applied to target detection applications by various academics, who

Received 25/September/2022; Accepted 08/December /2022

ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.23.2.11

had put out many great algorithms as well. The single-stage detection algorithm that is based upon region proposal and the 2-stage detection algorithm that has been based upon regression may be classified essentially to 2 groups.

B.1 Most important two-stage object detection algorithms

The most common two-stage detectors are those that (a) produce ROI using a region proposal network in the initial stage then (b) send region proposals farther along pipeline for bounding-box regression and object classification. These models have the maximum rates of accuracy, however, they are often slower, as is illustrated below in more detail.

B.1.1 RCNN and SPPNet (2014)

R-CNN, [2] algorithm, which represents the first practical model of target detection depending on CNNs, was put forth by Girshick in 2014. About 2000 region proposals for every one of the images that needs to be detected are initially extracted by the model using the selective search, as seen in *Fig. 2*. The retrieved image features are after that fed into SVM classifier for classification after being scaled uniformly to a feature vector of fixed-length. Lastly, a model of linear regression has been trained in order to carry out bounding box regression process. R-CNN does significantly enhance accuracy when put to comparison with the conventional detection approach, yet it requires a lot of calculations and does it inefficiently. Second, converting the region proposal to fixed-length feature vector directly might distort the objects.

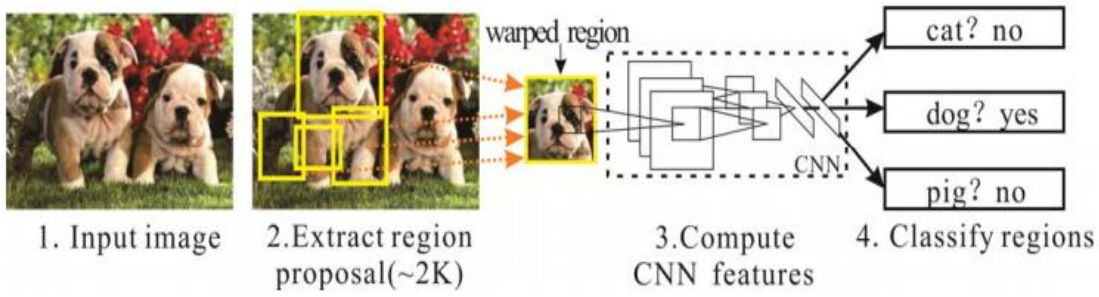


FIG. 2. R-CNN ARCHITECTURE.

The FCL imposes a fixed-size limitation on the input data that demands a fixed-length vector because the current CNNs demand that the input data be the same size. The accuracy of the classification might be impacted by the loss of image data caused by this artificial procedure. These problems with pattern recognition have been resolved by a novel CNN structure referred to as SPP-net, which adds a layer of SPP on top of the final convolution layer, as can be seen in *Fig. 3*. The FCLs (or other classifiers) are then fed with the fixed-length outputs produced by the SPP layer, which pools the features. With the SPP-net, CNN is able to accept inputs of any scale, enhancing the model's scale invariance, suppressing overfitting, and facilitating the extraction of local characteristics from the data at multiple sizes. The SPP-net is implemented by training each entire epoch on a single network and switching from one network size (224 × 224) to another (180 × 180). Following that, for the following complete epoch, the network size should be changed to the alternate (while keeping all weights). As a result, the majority of fixed-size images are trained on a single network, while images of various sizes are trained on a different network. Under such network switching, multiple networks' weights cannot be shared [8].

Received 25/September/2022; Accepted 08/December /2022
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.23.2.11

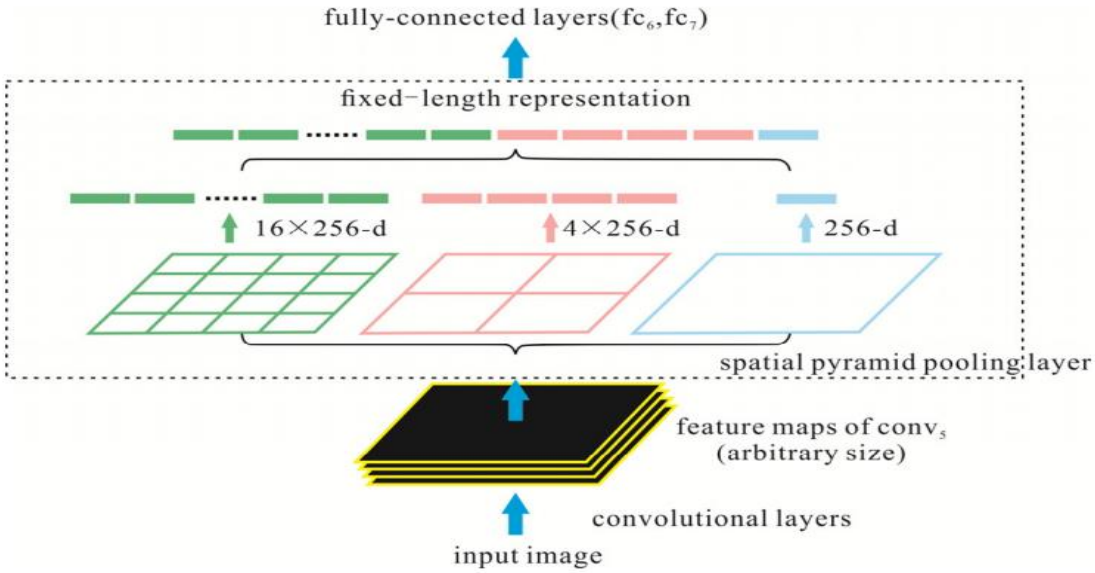


FIG. 3. SPP-NET ARCHITECTURE.

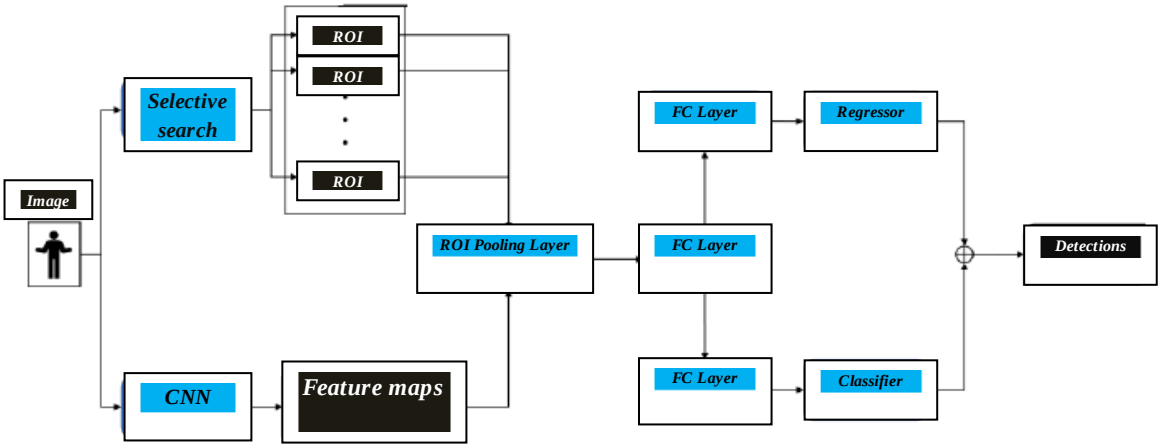


FIG. 4. (A) FAST R-CNN.

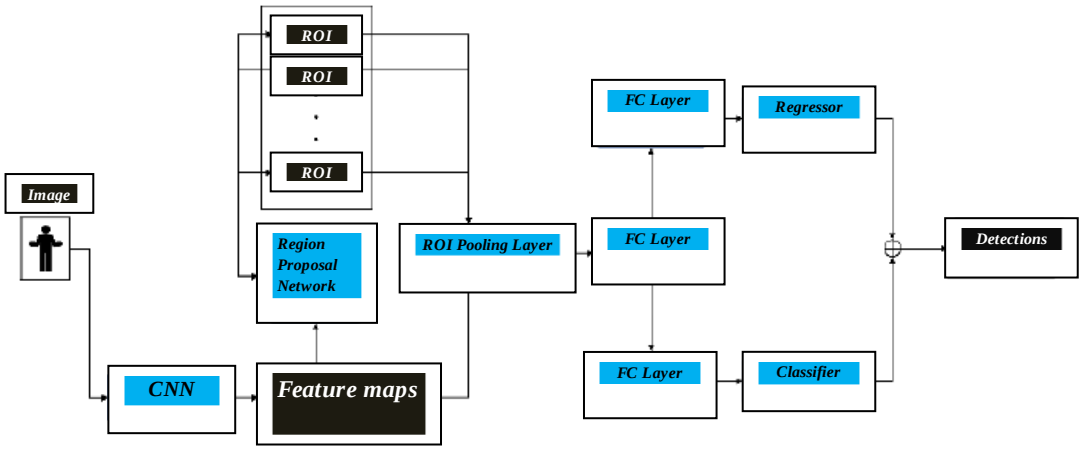


FIG. 4. (B) FASTER R-CNN.

Received 25/September/2022; Accepted 08/December /2022

 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.23.2.11

B.1.2 Fast RCNN and faster RCNN (2015)

Due to the fact that each ROI is fed to the CNN separately, the fundamental disadvantage of R-CNN as previously described is that it has an extremely long inference time. Girshick suggested a faster version of the algorithm referred to as Fast R-CNN that, rather than sending each ROI to the CNN individually, extracts feature maps of the whole input image simultaneously. The process of the ROI proposal is still used, yet a new ROI pooling layer called max pooling for each cell is added. Fully Connected (FC) layers use the output regarding such layer for predicting the object class and bounding box. This algorithm's flow is depicted in *Fig.4a*. With the addition of the RPN, which predicts ROIs using sliding window and anchor boxes, Faster R-CNN substantially enhanced the performance of this technique. This approach has the advantage that it could be trained for predicting bounding boxes which are more accurate, which leads to fewer predictions of low quality and a shorter inference time for each image. *Fig. 4b* depicts the algorithm's flow [9].

B.1.3 Mask R-CNN (2017)

With the use of very small datasets, Mask Region Based Convolutional Neural Network (Mask-RCNN) is an approach for highly accurate object detection. A pre-trained model depending on a Microsoft COCO data-set has been utilized as network which had likely been trained already for distinguishing basic features, and the model is adjusted additionally for minimizing validation loss in problem to combat overfitting and enhance generalization. The segmentation method known as Mask R-CNN may offer pixel-level boundaries for each recognized object. As demonstrated in *Fig. 5*, first, the feature map of the complete image is recovered utilizing Res Net-101 architecture as a convolutional backbone, giving Mask R-CNN a new capacity to separate objects in addition to detection and classification. The created feature map is next examined by a Region Proposal Network (RPN), which after that suggests potential options for object bounding boxes. A quantization-free layer referred to as ROI Align is used to address Faster R-problem CNNs with the pixel-to-pixel misalignment between network outputs and inputs while maintaining spatial locations. The network could classify objects and recognize bounding boxes after employing fully connected (FC) layers to correct the bounding-box candidates' misalignment problem. In addition, a convolutional layer unit predicts masks that are applied independently to each RoI [10].

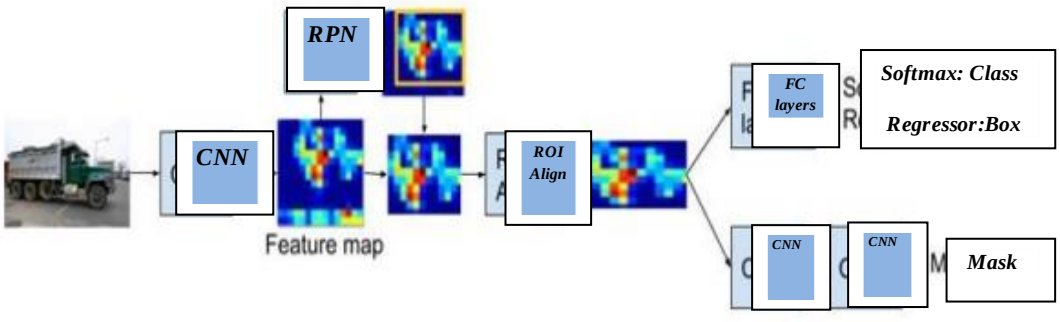


FIG. 5. THE ARCHITECTURE OF MASK R-CNN.

B.1.4 G-RCNN (2021)

For multi-object detection, a new deep CNN model called G-RCNN was introduced. This is a better iteration of the popular Fast RCNN and Faster RCNN that performs better. In comparison with the Faster RCNN and Fast RCNN, G-RCNN directly accepts video as input

Received 25/September/2022; Accepted 08/December /2022
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.23.2.11

and takes both temporal and spatial information into account. *Fig. 6*: Incorporating granulated layers employing spatio-temporal information within deep CNN architecture allows better object (s) localization (RoIs). The detection accuracy is greatly improved when the classification task is limited to objects found in ROIs. G-RCNN is faster than Fast RCNN and similar to the Faster RCNN in the terms of the speed [11].

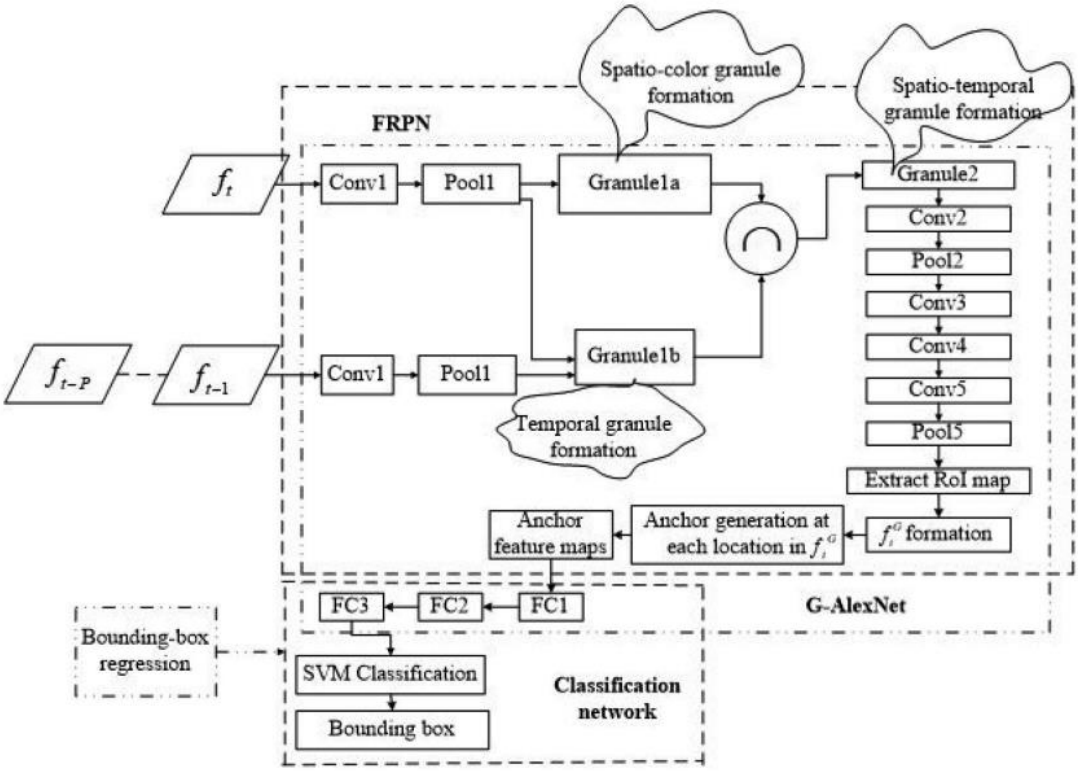


FIG. 6. G-RCNN ARCHITECTURE.

B.2 Most important one-stage object detection algorithms

The aforementioned object detection networks are two-stage networks; in the first stage, a region with a likely object is generated using a selective search proposal similar to the object proposal approach, and in the second stage, the proposed area is classified and the object is located. Debugging and network optimization are challenging with a two-stage network. Humans could often classify images in general terms at a glance when they observe how images are recognized and objects are located. However, testing and training times grow longer. That is, NNs must be able to recognize objects in real time. Due to the poor performance of networks like two-stage networks, one-stage networks have lately been deployed, as will be discussed below:

B.2.1 YOLO (2015)

YOLO was employed in numerous applications where object detection is necessary. The new network design, which Joseph Redmon suggested in 2015 and called YOLO, stands for You Only Look Once. This algorithm offers an alternative method for converting the object detection problem into a regression problem when put to comparison with the architecture of R-CNN family. It is able to immediately locate the target's classification category and bounding boxes at multiple locations across an input image. A convolutional neural network (CNN) called YOLO integrates the prediction of multiple bounding box locations and categories into a single stage. Yolo immediately selects the complete image training model

Received 25/September/2022; Accepted 08/December /2022
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.23.2.11

rather than the sliding window or the process of extracting the proposal to train the network. In *Fig. 7*, the architecture is displayed [12].

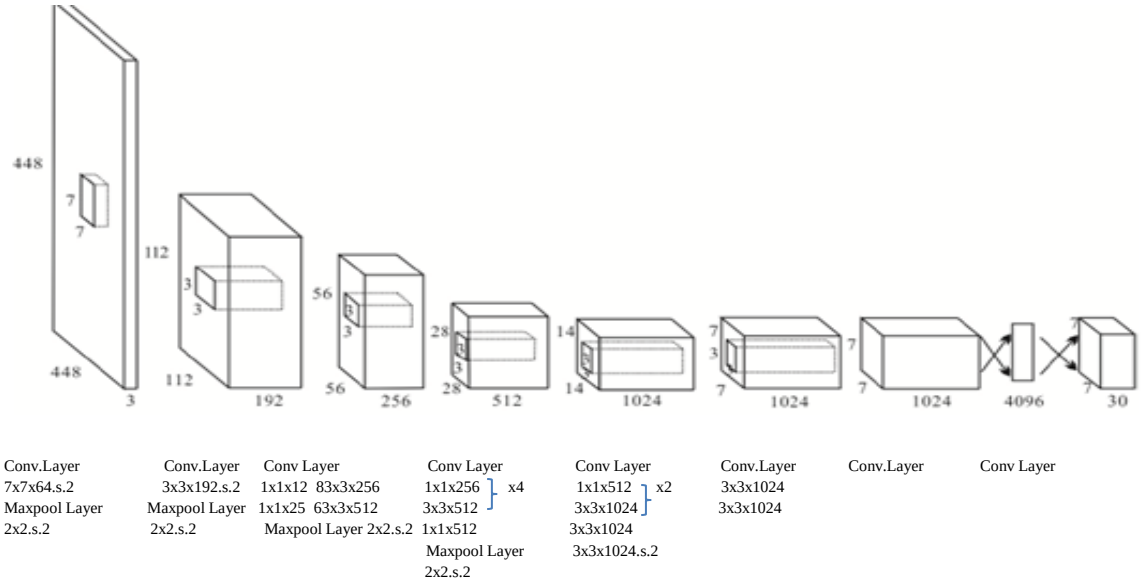


FIG. 7. THE YOLO ARCHITECTURE [2].

B.2.2 SSD (2016)

In comparison to solid-state drive (SSD), other models of object detection like YOLO or Faster R-CNN operate at a considerably slower rate, making them a far more practical object detection technique. Researchers came to the conclusion that instead of changing an existing model, they might have to find a fundamentally different model of object detection, which led to SSD model development. Yet, any notable increase in the speed by such modifications had just led to a reduction in the accuracy of the detection. The creation of the bounding boxes and feature map extraction, sometimes referred to as default bounding boxes, are critical components of its design. The network calculates loss by comparing the predicted classes' offsets and default bounding boxes with the values of the ground truth for training samples, applying a new filter for each iteration. All parameters are changed with the use of back-propagation algorithm and estimated loss value. As illustrated in *Fig. 8*, SSD can do this by learning the best possible filter structures which could precisely detect the object's features and generalize the provided training samples for minimizing the loss value. This leads to high accuracy throughout the assessment phase [13].

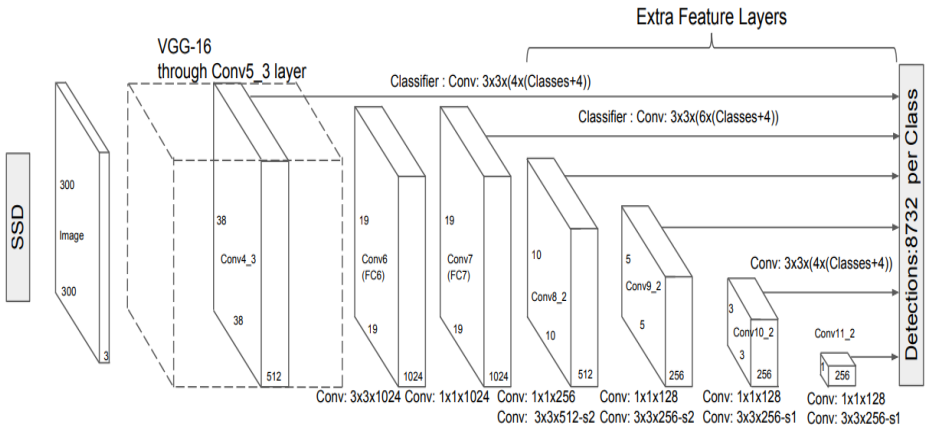


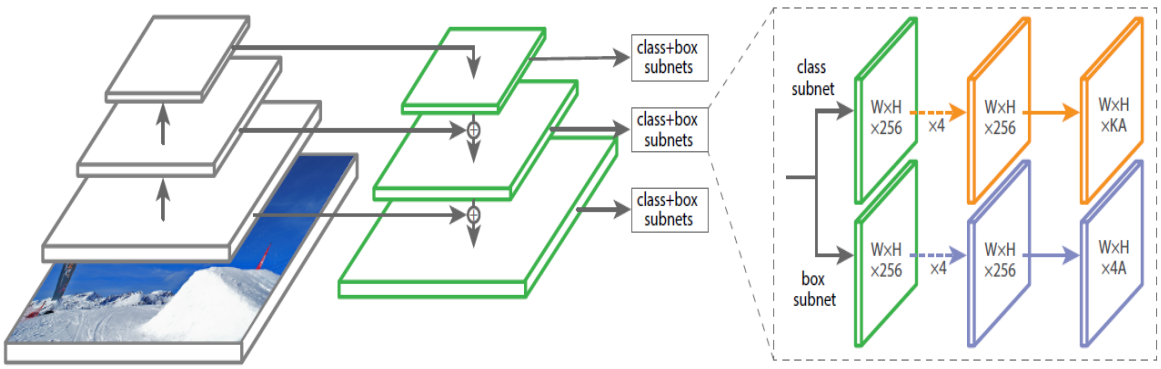
FIG. 8. SSD ARCHITECTURE.

Received 25/September/2022; Accepted 08/December /2022
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.23.2.11

B.2.3 Retina Net (2017)

Retina Net is a single, unified network that consists of a backbone network and 2 task-specific sub networks. The backbone, which is an off-the-self CNN, performs the calculation of a convolutional feature map over the entire input image. The backbone output is subjected to convolutional object classification in the first sub-net and convolutional bounding box regression in the second. *Fig. 9* depicts both sub networks, which have a straightforward design that suggest particularly for 1-stage, dense detection and are a part of Retina Net [14].



a) ResNet b) feature pyramid net c) class subnet (top) d) box subnet (bottom)

FIG. 9. RETINA NET ARCHITECTURE.

B.2.4 YOLOv3 (2018)

The next disadvantages prevent the base YOLO from developing as quickly as it could. Base YOLO has a solid edge in terms of acceptable speed and accuracy:

1- The very close proximity of two objects is a challenge for the base YOLO. In this condition, the detector might just be able to predict one object, which lowers the inference rate of the detector.

2- Although every one of the grid cells predicts two bounding boxes, only one type of objects is represented by the predictions. Therefore, if 2 objects are found in same grid cell, it cannot produce the correct results.

3- The fully-connected layer (FCL) used by the base YOLO to output the predictions must have the same dimension as its inputs.

The improved YOLO, YOLO3, was suggested in the year 2018, and several new concepts depending on YOLO2 have been incorporated. *Fig. 10* illustrates how the YOLO3 deepens the network structure by using 53 convolution layers (Darknet-53) as opposed to Darknet-19, which additionally inserts residual block into the network. For multiple label predictions, the logistic function is introduced in place
of the "softmax" function. The multiscale prediction of YOLO3, which enhances the algorithm's capacity for predicting small objects, is also a noteworthy accomplishment. For the purpose of increasing the accuracy and speed of the process of object detection, YOLO4 builds on the foundation of YOLO3 and incorporates certain unique technology, like cross stage partial connections, cross mini batch normalization, and weighted residual connections, among others [15].

Received 25/September/2022; Accepted 08/December /2022

ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.23.2.11

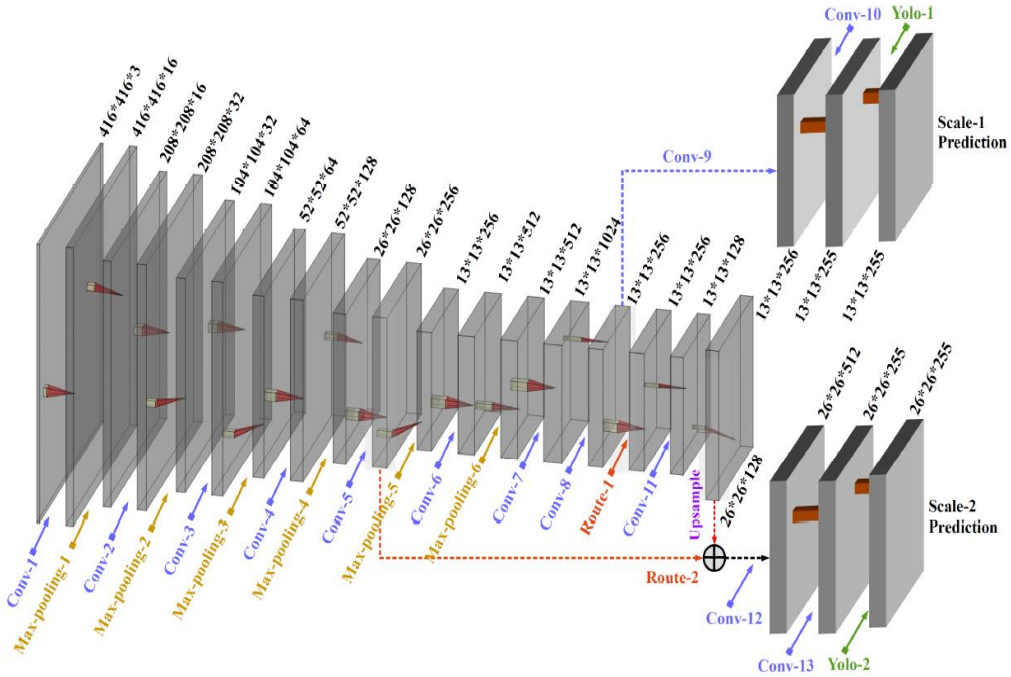


FIG. 10. NETWORK STRUCTURE OF TINY YOLO3. IT INCLUDES 6 MAX-POOLING LAYERS, 13 CONVOLUTION LAYERS, 1 UPSAMPLING LAYER, 2 ROUTE LAYERS, AND 2 YOLO LAYERS.

B.2.5 YOLOv4 (2020)

Using the basis of Joseph Redmon's theoretical concepts, Alexey Bochkovsky, a Russian engineer and researcher who constructed the Darknet framework and the first three iterations of the YOLO architecture on C, collaborated with Hon-Yuan and Chien Yao to publish YOLOv4 in April 2020 [13]. The single-stage detector series includes the YOLOv4 detector. The detector design, shown in *Fig. 11*, which consists of a neck, backbone, and head, is one of the fundamental characteristics that set YOLOv4 apart from preceding versions. Using CSPDarknet53 serves as backbone. The residual connections in this CNN prevent gradient vanishing and allow data to pass from initial layer to final layer.

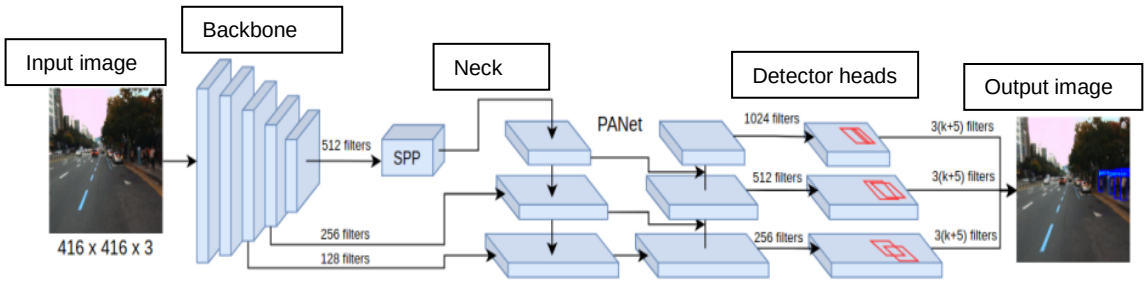


FIG. 11. YOLOV4 ARCHITECTURE.

In order to create a new feature extractor backbone referred to as CSPDarknet53, YOLOv4's upgraded implementation employs Darknet's Cross Stage Partial Network (CSPNet). Depending on modified DenseNet, the convolution architecture was developed. By using dense blocks, it moves a copy of feature map from base layer to the following one. The shrinking gradient vanishing problems, boosted backpropagation, elimination of the

Received 25/September/2022; Accepted 08/December /2022
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.23.2.11

computational bottleneck, and enhanced learning are benefits of adopting DenseNet. PANet path aggregation and Spatial Pyramid Pooling (SPP) layer make up the neck. For increasing receptive field and shorting out crucial characteristics from backbone, feature aggregation is done using the SPP layer and PANet path aggregation. Additionally, YOLO layer makes up the head. The image is fed to path aggregation network PANet for fusion after being fed to CSPDarknet53 for feature extraction. Comparable to YOLOv4 and YOLOv3, which also utilizes bag of specials and bag of freebies to enhance algorithm performance, YOLO layer then generates the results [16].

B.2.6 YOLO v5 (2021)

On the other hand, as illustrated in *Fig. 12*, YOLOv5 differs from earlier releases. Rather than Darknet, PyTorch is used. It utilizes the CSPDarknet-53 as backbone. This backbone eliminates the redundant gradient information seen in large backbones and incorporates gradient change into feature maps, and that improves accuracy, speeds up inference, and shrinks the size of the model through reducing the number of parameters. It boosts the information flow by using the path aggregation network (PANet) as a neck. The new feature pyramid network (FPN) that PANet uses has both bottom-up layer and top-down layer. Which enhances how low-level features in the model are propagated. The precision of the object's localization is increased because to PANet's improved localization in lower layers. Additionally, YOLOv5 has the same head as YOLOv3 and YOLOv4, which results in 3 distinct feature map outputs for the multi-scale predictions. Additionally, it improves the model's prediction for predicting objects of all sizes accurately [17].

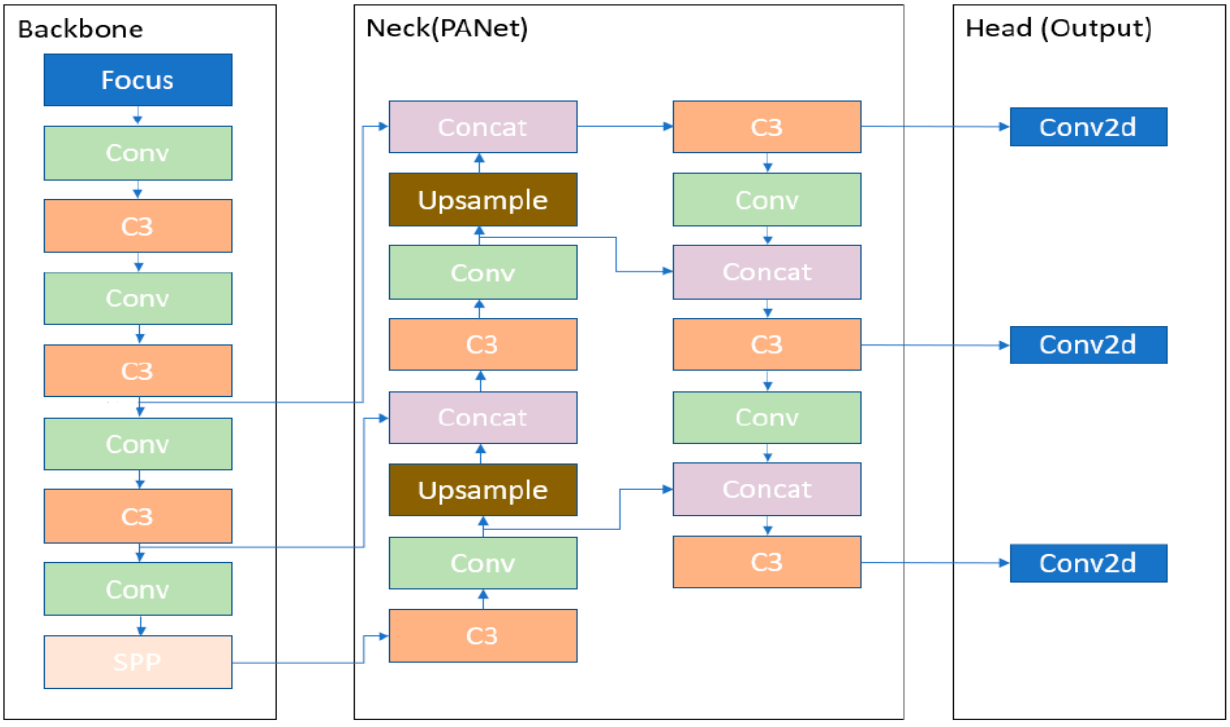


FIG. 12. YOLOV5 ARCHITECTURE.

B.2.7 YOLO v7

YOLO v7, a new YOLO model, is anticipated to become the next industry standard for object detection [18]. The number of computations, number of parameters, and computational density regarding a model are the main considerations in the construction of an effective YOLO v7 architecture. The next significant advancement in architecture search is referred to

Received 25/September/2022; Accepted 08/December /2022

ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.23.2.11

as YOLO v7. Researchers from YOLOv7 examined how re-parameterized convolution must be coupled with various networks using gradient flow propagation paths. In YOLOv7 architecture, the lead head is in charge of producing the output, and the auxiliary head is in charge of assisting in training. To create coarse-to-fine hierarchical labels for auxiliary head and lead head learning, respectively, YOLOv7 uses lead head prediction as guidance. *Fig. 13* depicts the two suggested deep supervision label assignment mechanisms.

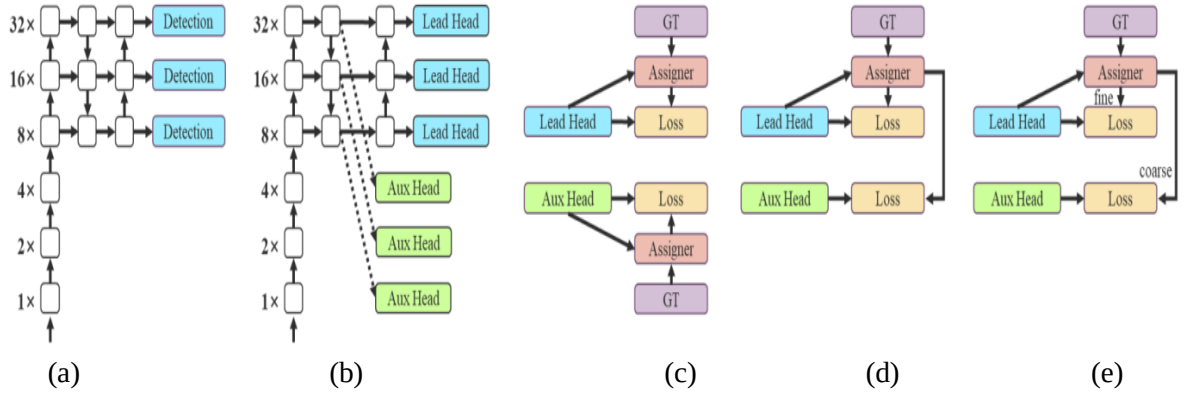


FIG. 13. COARSE FOR THE AUXILIARY AND FINE FOR THE LEAD HEAD LABEL ASSIGNER. COMPARED TO THE NORMAL MODEL (A), SCHEMA IN (B) HAS AN AUXILIARY HEAD. DIFFERENT FROM A TYPICAL INDEPENDENT LABEL ASSIGNER (C), THE PROPOSED (D) LEAD HEAD GUIDED LABEL ASSIGNER AND (E) COARSE-TO-FINE LEAD HEAD GUIDED LABEL ASSIGNER.

The conclusions drawn from the **YOLO v7** were that by controlling the shortest longest gradient path, a deeper network has the ability of learning and converging effectively.

IV. CONCLUSIONS

One-stage approach, which prioritizes inference speeds, is ultimately one of the two primary state-of-art approaches utilized for object detection. The classes and bounding boxes for the entire image are predicted in one-stage detector models; the ROI is not selected. The fundamental benefit of single-stage algorithms is that they're typically faster compared to multi-stage detectors and structurally simpler, making them faster than 2-stage detectors. However, two-stage algorithms have benefits in accuracy.

Object detection can be defined as one of the most fundamental and difficult issues in computer vision, and it had attracted a lot of attention lately. Although DL-based detection algorithms were majorly used in various industries, there are still some problems that need to be investigated:

1) Decrease dependence on data.
2) Achieving effective detection of small objects.
3) Achieving multi-category object detection.
4) Two-stage object detectors locate a ROI and classify the region using this cropped region. Yet, since cropping is a non-differentiable process, this type of the multi-stage detectors are typically not trainable from end to end.

Received 25/September/2022; Accepted 08/December /2022

ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.23.2.11

5) One-stage object detectors favor inference speed and are extremely quick, however, they struggle to identify groups of small objects or objects with unusual shapes.

REFERENCES

[1] K. Bayoudh, R. Knani, F. Hamdaoui, and A. Mtibaa, "A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets," Vis. Comput., vol. 38, no. 8, pp. 2939–2970, Aug. 2022.

[2] J. Deng, X. Xuan, W. Wang, Z. Li, H. Yao, and Z. Wang, "A review of research on object detection based on deep learning," in Journal of Physics: Conference Series, vol. 1684, no. 1, p. 12028, Mar.2020.

[3] C. Szegedy, A. Toshev, and D. Erhan, "Deep Neural Networks for Object Detection", Conference Paper, PP. 1-10, Feb 2020.

[4] J. Zhu, Z. Wang, S. Wang, and S. Chen, "Moving object detection based on background compensation and deep learning," Symmetry (Basel)., vol. 12, no. 12, pp. 1–17, Dec. 2020.

[5] S. K. Pal, A. Pramanik, J. Maiti, and P. Mitra, "Deep learning in multi-object detection and tracking: state of the art", Applied Intelligence, vol. 51, no. 9, pp. 1-31, Feb. 2021.

[6] M. Priyanka, K. Lavanya, K. C. Sai, and K. Rohit, "by Deep Learning Methods", Conference paper, pp. 1-16, Jul. 2022.

[7] M. Li, H. Zhu, H. Chen, L. Xue, and T. Gao, "Research on Object Detection Algorithm Based on Deep Learning," J. Phys. Conf. Ser., vol. 1995, no. 1, pp. 1-6, Jul. 2021.

[8] K. He, X. Zhang, S. Ren, and S. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition". IEEE Transactions on Pattern Analysis & Machine Intelligence, vol.4, PP. 1-14, Apr 2015.

[9] F. Dumitrescu, C.-A. Boiangiu, and M.-L. Voncilă, "Fast and Robust People Detection in RGB Images," Appl. Sci., vol. 12, no. 3, p. 1225, Jan. 2022.

[10] H. Raoofi and A. Motamedi, "Mask r-cnn deep learning-based approach to detect construction machinery on jobsites," Proc. 37th Int. Symp. Autom. Robot. Constr. ISARC 2020 From Demonstr. to Pract. Use - To New Stage Constr. Robot, no. Isarc, pp. 1122–1127, Nov. 2020.

[11] A. Pramanik, S. K. Pal, J. Maiti, and P. Mitra, "Granulated RCNN and Multi-Class Deep SORT for Multi-Object Detection and Tracking," IEEE Trans. Emerg. Top. Comput. Intell., vol. 6, no. 1, pp. 171–181, Feb.2022.

[12] D. Thuan, "Evolution of Yolo Algorithm and Yolov5: the State-of-the-Art Object Detection Algorithm," p. 61, Jan. 2021.

[13] U. Alganci, M. Soydas, and E. Sertel, "Comparative research on deep learning approaches for airplane detection from very high-resolution satellite images," Remote Sens., vol. 12, no. 3, pp.1-28, Feb. 2020.

[14] T. Lin, P. Goyal, R. Girshick, K. He, P. Dollar, "Focal Loss for Dense Object Detection", arXiv:1708.02002v2 [cs.CV], vol.2, pp. 1-10, Feb 2018.

[15] T. Li, Y. Ma, and T. Endoh, "A Systematic Study of Tiny YOLO3 Inference: Toward Compact Brainware Processor with Less Memory and Logic Gate," IEEE Access, vol. 8, pp. 142931–142955, Aug. 2020.

[16] K. Roszyk, M. R. Nowicki, and P. Skrzypczynski, "Adopting the YOLOv4 Architecture for Low-Latency Multispectral Pedestrian Detection in Autonomous Driving", journal/sensors, pp. 1-21, Jan 2022.

[17] U. Nepal and H. Eslamiat, "Comparing YOLOv3, YOLOv4 and YOLOv5 for Autonomous Landing Spot Detection in Faulty UAVs," Sensors, vol. 22, no. 2, pp. 1-15, Jan. 2022.

[18] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," arXiv Prepr. arXiv2207.02696, pp. 1-13, Jul. 2022.

Received 25/September/2022; Accepted 08/December /2022

 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352