International Journal of Research in Engineering, Science and Management
Volume 4, Issue 7, July 2021
https://www.ijresm.com | ISSN (Online): 2581-5792

Real-Time Object Detection with Pre-eminent Speed and Precision using YOLOv4

Shraddha Sanjeev Pattanshetti[1*], Shabana Imam Nivade[2]
[1,2]*Student, Department of Information Technology, D.K.T.E's Society Textile and Engineering Institute, Ichalkaranji, India*

***Abstract*: A computer vision technique that has fascinated the world with its outstanding ability to localize and identify objects is Object recognition. It draws bounding boxes around the recognized objects and accurately labels them. Object detection is more intricate than classification as it not only recognizes the object but also the location of the object in the image. A widely known algorithm for precise and quick detection is YOLO. YOLO (You only look once) is an open-source and reliable real-time object recognition algorithm that can identify multiple objects in a single frame. Furthermore, it recognizes objects more rapidly and precisely than other recognition systems. It is one of the best and adaptable computer vision algorithms because it can process 45 frames per second and can estimate up to 9000 and more seen and unseen classes of objects. It is also seen that YOLO works faster than RCNN, due to its elementary architecture. Antithetical to R-CNN algorithms that use regions to circumscribe the entities in images, YOLO instead applies neural network to the complete image to anticipate bounding boxes and their probabilities. YOLOv4 is finer than YOLOv3 in terms of Adequate Speed and Accuracy of Object Recognition.**

***Keywords*: artificial intelligence, computer vision, deep learning, image processing, machine learning, neural network, object detection, pattern recognition, video processing.**

1. Introduction

YOLO is an acronym for You Only Look Once that is object detection model that belongs to one stage detectors. YOLO was implemented to ameliorate accuracy and speed allowing them to not only work as recommendation systems but also for stand-alone process management and human input reduction. Other object detection methods have drawbacks that require a greater number of GPUs for training unlike YOLO. There are endless applications where YOLOv4 can be used like Mask Detection, Signature Detection, Table Detection, Object Counter and so on.

YOLOv4 overtakes existing algorithms in terms Performance (FPS) and Speed (AP). Hence the objective for innovating YOLOv4 was to develop an object detector algorithm that can run rapidly in production system and can be optimized to work smoothly in parallel computations.

There are versions prior to the YOLOv4. There were various problems in those algorithms that made scientist develop YOLOv4 with great enhancements and improvements. Firstly, YOLOv1 was developed that had 26 neural network layers and used 2 fully connected layers. Moreover, it had 24 convolutional neural layers. The main reason that YOLOv1 became less popular was its incapability of detecting tiny or small objects. In order to overcome this serious problem YOLOv2 or YOLO9000 was developed. It had altogether 30 neural network layers. Furthermore, batch normalization was introduced after every convolutional neural network. Also, a new concept was added named 'Anchor Boxes'. Unlike YOLOv1, fully connected neural layers are present in YOLOv2. Images ranging from 320 to 608 were taken for training. Multiple labels can be given to same image but it can encounter multiclass problem also called WordTree concept. One of the two label appears (Either parent or child) not both. Later, scientist discovered YOLOv3 that was capable to solve the problems that YOLOv2 could not. It has 106 neural network layers. It works fine when it comes to tiny objects. Some changes have been made in error functions. Multiclass problems have turned into multilabel problem. Since YOLOv3 has 9 anchor boxes, 3 per scale, so bounding boxes can be predicted more compared to YOLOv1 and YOLOv2.

There are some major improvements in YOLOv4 when compared to prior YOLO versions. Firstly, YOLOv4 is very competent and attested model that authenticates everyone to use 1080 Titanium or 2080 Titanium GPU for training models with ultra-speed and accuracy. Also, techniques like bag-of-freebies and bag-of-specials are bought into exercise to improvise the accuracy or model during training and post processing. Usage of modified state-of-art functions including Cross-iteration Batch Normalization (CBN), Path Aggregation Network (PAN) for single GPU Training in YOLOv4 has a greater advantage over other YOLO versions. Furthermore, there are few enhancements in YOLOv4 when compared to other existing object detection models. YOLOv4 has attained a Speed (AP) value of 43.5% on Microsoft COCO Dataset and hence considered one of the best models in terms of speed and accuracy. Also, in terms of Speed (AP) and Performance (FPS), YOLOv4 is double as rapid as EfficientDet. Compared to YOLOv3 and other models available, Average Precision and Frames per Second have increased by 10% and 12% respectively.

*Corresponding author: shraddhapattanshetti161998@gmail.com

Along with advancements, YOLO includes some limitations. YOLO has less recall factor compared to Faster R-CNN and has excess localization error when compared to Faster R-CNN. Detecting close objects in YOLO is difficult because each grid offers only 2 bounding boxes. When it comes to detecting smaller objects, it becomes arduous task for YOLO. Architecture of YOLO cannot attain state-of-art-accuracy. It becomes difficult for algorithm to localize objects compared to Faster R-CNN.

A. Literature Review

You Only Look Once (YOLO) is computer vision algorithm that has advantage over other deep learning object detection models in terms of speed and performance.

YOLOv4: Optimal Speed and Accuracy for object detection, this article clearly gives us overview of what YOLOv4 comprises. More specifically, it mentions in detail information about Bag-of- Freebies and Bag-of-Specials. Furthermore, it gives in depth knowledge of YOLOv4 design that includes Backbone, Neck and head of YOLOv4 (Anka, 2020).

Breaking Down YOLOv4 article gives a generalized overview of what Object Detection does. Later it talks about object detection anatomy by explaining architecture of two staged detector in deep. Also, it gives brief knowledge of prior YOLO versions. Moreover, it explains in detail about CSPResNext50, CSPDarknet53 and EficientNet-B3 as backbone of YOLOv4. Later, it mentions various options that can be considered as Neck of YOLOv4 which includes FPN, PAN, NAS-FPN, BiFPN, ASPP, SFAM (Solawetz, 2020).

A Gentle Introduction to YOLOv4 for Object detection in Ubuntu 20.04 specifically talks about object detection and its need. Also, gives information about Scientist that created YOLO. Furthermore, it gives information regarding darknet and steps for installing YOLO on Ubuntu 20.04 (Joseph, 2020)

Research Paper on Vehicle Object Detection Algorithm Based on Improved YOLOv3 Algorithm notifies us how YOLOv3 lacks in terms of accuracy, speed and unsatisfactory results during vehicle detection and thus introduces an algorithm named F-YOLOv3 that shows improved results. It gives in depth information regarding F-YOLOv3 network structure, Improvements of Anchor Box Dimension Clustering Algorithm and Improved Multi-scale Prediction (Liu & Zhang, 2020, p. 10)

IEEE paper "You only Look Once: Unified, Real-Time Object Detection" mainly talks about new object detection approach and also compares how YOLO is better that other Deep learning-based object detection like R-CNN. It mentions limitations of YOLO (Redmon, Divvala, Girshick, & Farhadi, 2016).

Article "Introduction to YOLOv4" specifies working of YOLOv4 and more specifically talks about advancements in YOLOv4 in comparison to prior YOLO models. In short, it talks about overview of YOLOv4 in terms of Speed, Precision and its network architecture (Tyagi, 2020) .

2. Working of YOLOv4 Algorithm

YOLOv4 is known for its upgradation in terms of AP and FPS. YOLOv4 prioritizes real-time object detection and training takes place on single CPU. YOLOv4 has obtained state-of-art results on COCO dataset with 43.5% speed (AP) at 65 Performance (FPS) on Tesla V100. This achievement is the result of combination of the features like DropBlock Regularization, Data Augmentation, Mish-Activation, Cross-Stage-Partial-connections (CSP), Self-adversarial-training (SAT), Weighted-Residual-Connections (WRC) and many more.

There are two types of models, one and two staged object detectors. In two stage detectors works in two parts that is first regions of importance are detected and then regions are classified to see if object is detected in that particular region. YOLOv4 being single staged object detector, works more accurate and faster than Two staged detectors like R-CNN, Fast R-CNN.

A. Object Detection Anatomy

Object detection Algorithms have generally 4 components that are mentioned below:

1) Input

Object detection is an aspect of computer vision and image processing wherein real-time identification and recognition of object takes place with help of image or video. Only after image or video is taken as input, backbone can be used for image processing in order to extract the features. Hence, Input is the first stage in from where the working of algorithm starts.

2) Backbone

Backbone as name suggests plays very cardinal role in object detection. It refers to feature extracting network that extracts features from the input provided. It gathers picture constituents or elements to form features at different aggregations. Accuracy of the detector completely depends on how precisely backbone extracts the features.

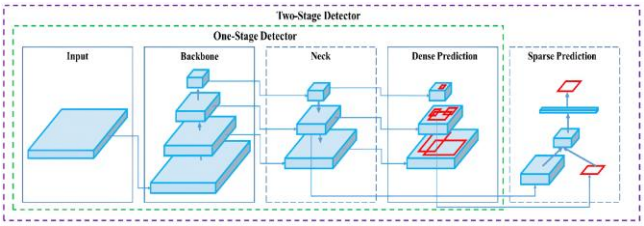

Fig. 1. Object Detection Architecture

3) Neck

Neck plays an essential part in gathering feature maps from different stages. In order to enhance the information to be inserted into head of the detector, adjoining feature maps received from top-down and bottom-up paths are combined element wise before giving it to head.

4) Head

After providing input in form of video or image, backbone selected extracts the features and then these features are combined and provided to head. Main operation performed is dense prediction. This is the stage where actual detection of

bounding boxes is done. It consists of coordinates of the bounding boxes that includes center, height, width, score of prediction and label.

B. YOLOv4 Architecture

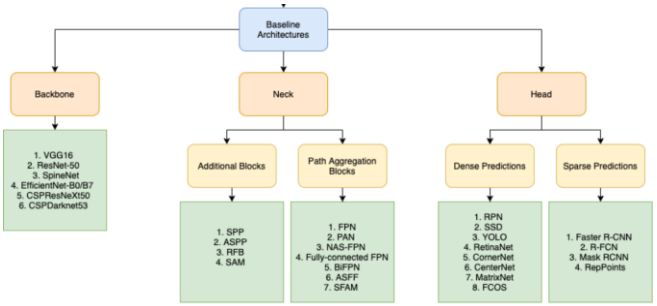

Fig. 2. Basic YOLOv4 Architecture

1) Backbone

YOLOv4 basically uses one of the three model as its backbone. Three feature extractor models include:

- CSPResNext50

  Both CSPResNext50 and CSPDarknet53 are DenseNet based models. It works similar to CSPDarknet53 that operates on CSPNet strategy. Considering COCO dataset, CSPDarknet53 is better in classifying objects than CSPResNext50. CSPResNext50 consists of 16 CNN layers with 425×425 receptive field and 20.6 M parameters while CSPDarknet53 consists of 29 CNN layers with 725×725 receptive field and 27.6 M parameters.

- CSPDarknet53

  Both CSPResNext50 and CSPDarknet53 are DenseNet based models. It works similar to CSPDarknet53 that operates on CSPNet strategy. Considering COCO dataset, CSPDarknet53 is better in classifying objects than CSPResNext50. CSPResNext50 consists of 16 CNN layers with 425×425 receptive field and 20.6 M parameters while CSPDarknet53 consists of 29 CNN layers with 725×725 receptive field and 27.6 M parameters.

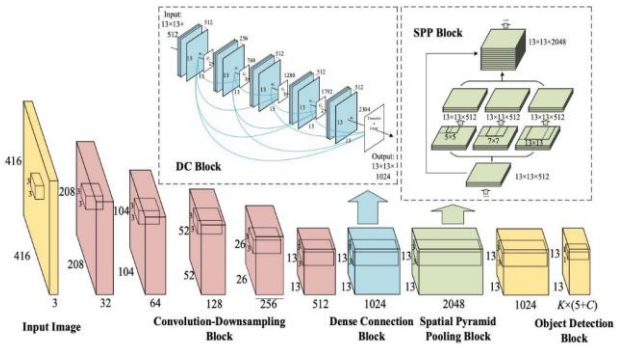

Fig. 3. CSPDarknet53

- EfficientNet-B3

  It is used as image Classification Model that is particularly used to attain state-of-art accuracy. It is generally utilized to restudy the Convolutional Neural Network scaling and is based on AutoML. AutoML mobile Framework was developed in order to develop a small sized network known as EfficientNet-B0. The Compounding Scaling as name suggests helps in scaling up AutoML baseline in order to gain Efficient-B1 to Efficient-B7.

2) Neck

This is the second stage where features are gathered that where initially formed in the backbone and later these features will be fed to head for detection. YOLOv4 has several options:

- FPN (Feature Pyramid Network)

  Earlier, detectors used to predict objects on pyramidal feature hierarchy extracted from backbone. In order to solve this major issue of effective representation and multi-scale feature processing, FPN was proposed that follows top down path to gather features of different scales.

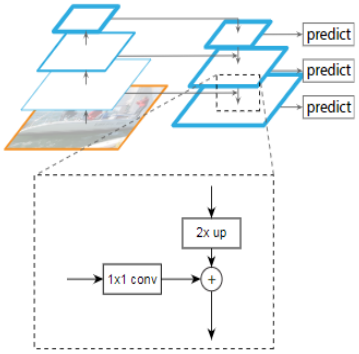

Fig. 4. FPN (Feature Pyramid Networks)

- PAN (Path Aggregation Networks)

  It is used as neck in YOLOv4 Algorithm and is used to improve the process of segmentation by maintaining semantic data or information precisely that helps in correct localization of picture elements for mask information.

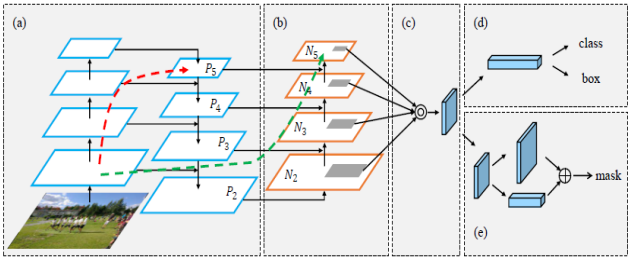

Fig. 5. PANet Architecture

- *Properties of PAN*

  1. Bottom-up Path Augmentation

     When image is passed through various layers of Neural network, spatial resolution of image reduces while complexity of the feature increases simultaneously. This results in complications in identification of pixel-level masks by high level features. In YOLOv3 FPN follows top down approach for combining multi-scale features preserving semantic localization information. Huge objects mask production becomes intricate and lengthy because the information needs to travel a greater number of layers.

On contrary, PANet includes an additional bottom up and top down route than FPN. This considerably reduces the complexity by usage of lateral connections from lower layers to upper. It has at most 10 years which makes it less complicate and lengthy.

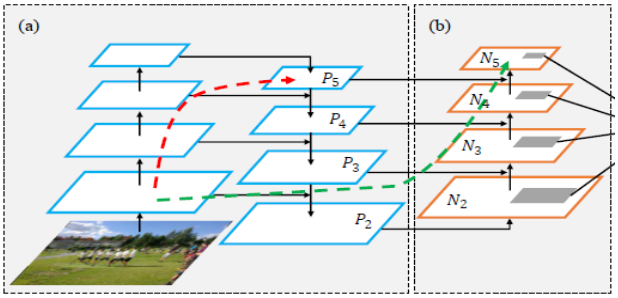

Fig. 6. Bottom-up Path Augmentation

2. Adaptive Feature Pooling

Earlier, techniques like Mask-RCNN made used of features from only one stage for prediction. RIO Align Pooling was used to uproot features from upper levels. At times result were predicted inaccurately. So, in order to overcome this problem PANet extracted features from all the layers which performs Align Pooling on every feature map to uproot features.

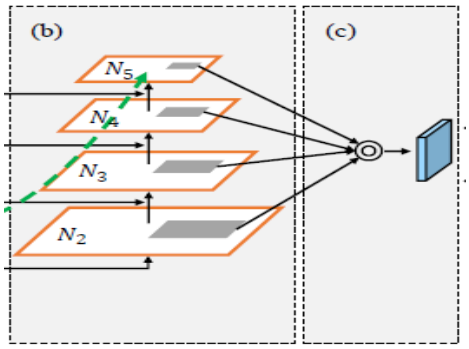

Fig. 7. Adaptive Feature Pooling

3. Fully-Connected Fusion

Fully Convolutional Network was used in Mask-RCNN to conserve spatial information and simultaneously number of parameters are also reduced. The model cannot learn how to use pixel location for prediction since the parameters are shared for all the spatial positions. Fully-connected layers are location sensitive and can adjust with different locations. PANet uses data from both layers for precise prediction.

- SPP (Spatial Pyramid Pooling)

SPP also known as Spatial Pyramid Pooling is used for obtaining both fine and coarse information. Sliding Kernel of dimension 1×1, 5×5, 9×9, 13×13 to which pool is applied. Maps of various kernel sizes are combined to get output. The main advantage to using SPP is to improve receptive field. Spatial Pyramid Pooling is used to create fixed size features regardless of feature map sizes. Advantages of SPP are it creates fixed result whatever might be input feature map size. Also, it makes use of multi-level spatial bins hence it is robust when it comes to object deformations. SPP is flexible in terms of input scales as it can extract pool features at variable scales.

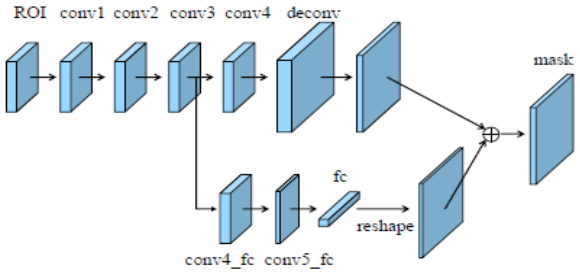

Fig. 7. Fully-Connected Fusion

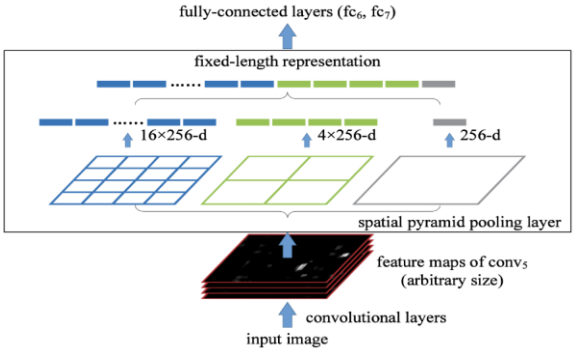

Fig. 8. Spatial Pyramid Pooling

3) Head

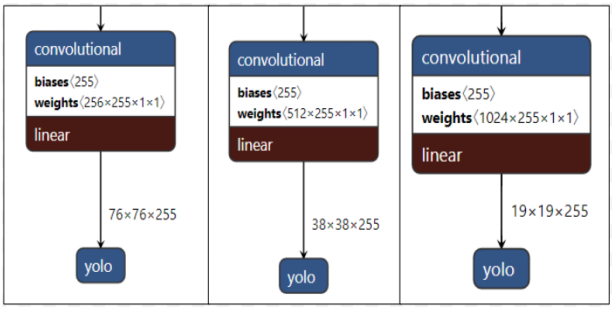

Fig. 9. YOLOv4 head applied at different scales

The main objective of the head in YOLOv4 is to perform prediction that includes classification and regression of bounding boxes. It uses YOLOv3 head. It provides information regarding coordinates of bounding boxes (x, y, h, w). It includes width, height, center and score of prediction with label. YOLOv4 head can be applied to every anchor box.

- Anchor Box

Anchor boxes are used for multiple objects of various sizes in a single frame whose center is positioned in the same cell. On contrary, grid was used to detect single object in a frame. If number of anchor boxes change then length of ground truth and prediction array also changes.

Consider a box in cell has 80 prediction classes i.e. [Pc, P1, P2... P80, X1, Y1, X2, Y2] which totals to 85, for these

9 anchor boxes will be used and will have array length of 85*9 = 765 predictions.

Example of anchor box plated around (0,0) of different scales.

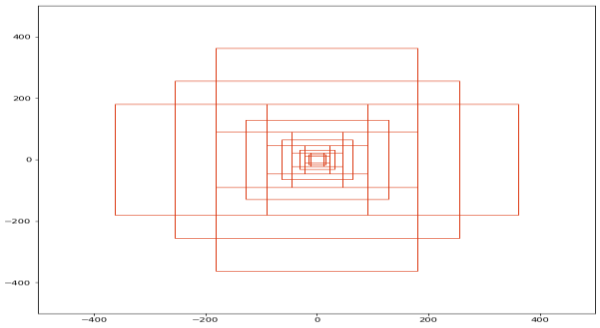

Fig. 10. Anchor Boxes

C. Bag of Freebies YOLOv4

Bag of Freebies can be termed as techniques that improvise cost related to training or strategy to ameliorate the accuracy of model. It increases the performance of model without compromising the latency at inference time and hence improvements are seen in data management and data augmentation. These methods enhance and augment the training dataset of model that are opened to situations that otherwise might have unsensed. Data Augmentation hence can be seen as a strategy to get utmost information from the dataset and would overall generalize the data or increase the robustness so that model can learn through a shallow range of situations.

D. Bag of Specials YOLOv4

Bag of specials consists of coherence and post-processing models due to which inference cost of the modules escalate by a fraction. However, this change helps the detector improve the accuracy and performance. Selection of the technique totally depends on architecture and other technical parameters but end outcome of refining is attained.

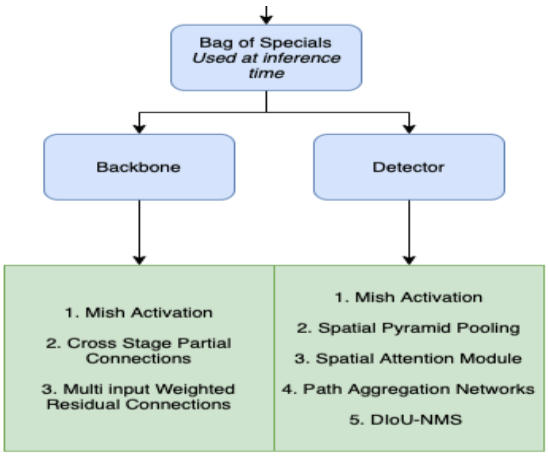

Fig. 11. Different methods present in Bag of Specials

3. Experimental Results – YOLOv4

In order to test YOLOv4 techniques, experiment was performed on COCO dataset. COCO dataset consists of 80 objects that emphasizes on a wide range of scenarios that is required for an object detector to predict accurate results. YOLOv4 performs a thorough ablation study. Ablation study can be termed as removal of features of the module/model in order to check the performance of model. Features addition or removal will define whether the network is improving or not. YOLOv4 finally attains state-of-art performance for object detection.

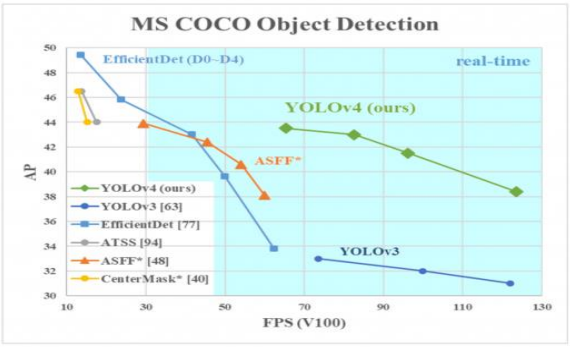

Fig. 12. Speed (AP) vs. Performance (FPS) of YOLOv4 and different models

4. Conclusion and Future Work

To conclude, YOLOv4 is one of the most efficient, flexible, robust, lightweight and easy to use algorithm for real time object detection. Practically, experiments have proved that YOLOv4 is better than R-CNN, Fast R-CNN, Faster R-CNN etc. Also, YOLOv4 is proved better in terms of speed (AP) and Performance (FPS) compared to previous YOLOv4 algorithms.

YOLOv5 recent version of YOLO can be used as unified real-time object detection because it is implemented in Ultralytics PyTorch Framework that can attain higher FPS. YOLOv5 is very easy to use, flexible and reliable when it comes to custom data in initial runs.

References

[1] A gentle introduction to YOLO V4 for object detection in Ubuntu 20.04. (2020, May 16). Robotics, Computer Vision, Embedded System, AI. https://robocademy.com/2020/05/01/a-gentle-introduction-to-yolo-v4-for-object-detection-in-ubuntu-20-04/#How_YOLO_work.

[2] Anka, A. (2020, July 16). YOLO V4: Optimal speed & accuracy for object detection. Medium. https://towardsdatascience.com/yolo-v4-optimal-speed-accuracy-for-object-detection-79896ed47b50#56ff.

[3] Breaking down YOLOv4. (2021, March 4). Roboflow Blog. https://blog.roboflow.com/a-thorough-breakdown-of-yolov4/.

[4] Chaudhari, S., Malkan, N., Momin, A., & Bonde, M. (2020). Yolo real time object detection. International Journal of Computer Trends and Technology, 68(6), 70-76.

[5] Liu, J., & Zhang, D. (2020). Research on vehicle object detection algorithm based on improved YOLOv3 algorithm. Journal of Physics: Conference Series, 1575, 012150.

[6] Lu, Y., Zhang, L., & Xie, W. (2020). YOLO-compact: An efficient YOLO network for single category real-time object detection. 2020 Chinese Control and Decision Conference (CCDC).

[7] Medium. (n.d.). Medium. https://medium.com/visionwizard/yolov4-version-2-bag-of-specials-fab.

[8] R, M. (2020, July 3). PANet: Path aggregation network in YOLOv4. Medium. https://medium.com/clique-org/panet-path-aggregation-network-in-yolov4-b1a6dd09d158.

[9] Redmon, J., & Farhadi, A. (2017). YOLO9000: Better, faster, stronger. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[10] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[11] Tyagi, N. (2020, 25). Introduction to YOLOv4. Analytics Steps | Learn AI, Machine Learning, Business Analytics, Startups & other Technologies. https://www.analyticssteps.com/blogs/introduction-yolov4.

[12] Wang, C., Mark Liao, H., Wu, Y., Chen, P., Hsieh, J., & Yeh, I. (2020). CSPNet: A new backbone that can enhance learning capability of CNN. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).

[13] YOLOv4 - Superior, faster & more accurate object detection. (n.d.). Artificial Intelligence & Augmented Reality Tutorials. https://www.augmentedstartups.com/blog/yolov4-superior-faster-more-accurate-object-detection.

[14] Zhang, X., Dong, X., Wei, Q., & Zhou, K, Real-time object detection algorithm based on improved YOLOv3. Journal of Electronic Imaging, 28(05), 1, 2019.