Observing the effects of Image Quality on Object Detection Using YOLOv5

Kshitij Malvankar[1, 2], Enda Fallon[1], Paul Conolly[2], Kieran Flanagan[2]

[1] Software Research Institute, Athlone Institute of Technology, Athlone, Ireland
kmalvankar@ait.ie, efallon@ait.ie

[2] The NPD Group, L.P, Athlone, Ireland
paul.connolly@npd.com kieran.flanagan@npd.com

***Abstract* - Document information extraction, which combines item classification with object localization within a scene, is a major difficulty in computer vision. With the advent of modern advances in deep learning, significant advancements in object detection have been made. Majority of research is focused on designing increasingly more complex object detection networks for improved accuracy, such as YOLOv5, SSD, R-CNN, Faster R-CNN, and other extended variants of these networks. This paper proposes to use the YOLOv5 algorithm to identify data in an invoice and also observe the effects of degraded image quality on the performance of the algorithm. The BRISQUE score is utilized to parameterize the quality of the image. The difference in performance under the same training conditions between three different variants of the YOLOv5 algorithm are also detailed.**

Keywords - Object Detecion, Faster R-CNN, YOLOv5, BRISQUE

I. INTRODUCTION

Data analytics has the potential to alter every facet of a retail organization, offering important information to departments ranging from logistics to administration, marketing, and sales. Retail analytics is the process by which retailers find, collect, and interpret data created across their departments. These procedures produce a set of actionable insights based on business trends, developing patterns, and performance metrics. An organization may use these insights to improve critical parts of a retail business such as consumer experience, logistics, administration, and even sales. Data analytics boosts competitiveness by helping organizations take data driven decisions.

Many businesses, particularly commercial entities, government agencies, medical institutions, and public security agencies, must issue a considerable number of invoices every day. These invoices include a considerable amount of valuable information such as the date and time of the purchase, the products bought and their prices and the payment information and analyzing this data can give any organization the insights necessary for improving its efficiency. This job can be highly manual and labor intensive, driving the cost up. Image recognition technology is often used to rapidly and reliably extract invoice information to tackle this challenge. Invoice information record accuracy can also be enhanced. Based on predetermined labels, an image detection problem predicts the label of an image. It selects a single object of interest in the image based on the assumption and attempts to cover a significant portion of the image. The detection duty entails not only determining the object's class, but also determining the extent of the object in the image.

A. Benefits

Using the Deep CNN has some benefits over traditional data entry methods:

A1. Cost Effective: This method allows an organization to save money on hiring workforce for manual data extraction. Employees can be redirected to other tasks, increasing the productivity of the organization.

A2. Error Reduction: Because of the many formats, extracting information from invoices is complicated. Human errors are also a significant issue, resulting in data loss and inaccuracy. This approach aids in the reduction of human error and ensures that the detection is accurate.

Section II of the paper details related work in the area of Object detection using the YOLOv5 algorithm. Section III comprises of the dataset design that was used for this paper while section IV details the methodology undertaken for the project. Section V details all the results that were observed after conduction the various experiments listed in section IV and section VI includes the conclusion and sets the pace for the proposed future work.

II. RELATED WORK

On the mainstream Microsoft Common Objects in Context (MS COCO) dataset [6], there exist two broad classifications of object detectors that continue to perform well [7]. In one-stage detection, YOLO [8], RetinaNet [9] techniques are often employed, while in two-stage detection, Faster R-CNN [4] or Mask R-CNN [10] methods

are frequently utilized. Mask R-CNN is a Faster R-CNN modification with an extra mask proposal branch for segmentation. YOLO uses a single neural network to predict bounding boxes and class probabilities from complete photos in a single evaluation. Because the whole detection pipeline is a single network, detection performance can be improved end-to-end. Faster R-CNN is a region-based technique that predicts detections based on local area attributes. A Region Proposal Network is used to localize this region (RPN). The proposal of the region based on the features obtained in the convolutional backbone form the first stage while the fully connected network for object classification and bounding box regression forms the second stage.

A. Single-Stage Object Detector

YOLO [8] and RetinaNet [9] are two well-known single-stage object identification models. Single-stage detectors often have modest computation needs and may be readily implemented on mobile devices. YOLO is a unified, real-time object identification system that reduces the object detection work to a single regression issue. YOLO predicts bounding boxes and class probabilities directly from complete pictures using a single neural network architecture. When compared to Faster R-CNN [4], YOLO delivers faster detection at the expense of accuracy.

B. Multi-Stage Object Detector

Region-based CNN (R-CNN) [3] is a type of object detection model that falls under the category of muti-stage detectors. Faster R-CNN [4] is a region-based technique for predicting detections using information from a suggested region. Faster R-CNN is a common two-stage detector that is based on region proposals. The first step provides a limited collection of candidate objects based on shared feature maps using a Region Pooling Network (RPN), which is categorized as foreground or background class. Hyperparameters are used to customize the size of each anchor. The ideas are then utilized to construct sub feature maps in the area of interest pooling layer (RoI pooling).

The sub feature maps are transformed to 4096-dimensional vectors before being input into fully linked layers. These layers are then combined to form a regression network that predicts bounding box offsets, while a classification network predicts the class label of each bounding box proposal. The network's backbone is the Feature Pyramid Network (FPN) [11]. FPN constructs an in-network feature pyramid from a single-scale input using a top-down design with lateral connections. Faster R-CNN with an FPN backbone takes RoI features from various layers of the feature pyramid according on their size, but the rest of the technique is the same as vanilla Resnet.

C. Applications of YOLOv5

YOLOv5 [1] has been used to detect heavy goods vehicles at rest zones, allowing for real-time parking place occupancy prediction [12]. The use of transfer learning was used to see if the front cabin and rear are sufficient attributes for recognizing big cargo vehicles. Experiments have demonstrated that, while the strategy is effective, it has to be improved in order to better accomplish this difficult task.

A method was proposed to detect smoking behavior based on the YOLOv5 algorithm [13]. Cigarettes being small targets in an input image, a K-means algorithm was used, and a small target detection layer was added to the YOLOv5 algorithm. The false detection rate on the self-created data set was 0%, and the AP was 92.3 percent, which is 6.7 percent higher than the YOLOv5s algorithm.

Sign language is used by people who have speech or hearing difficulty. However, it often happens that these signs and gestures are not understood or recognized by the general population. Due to it being lightweight, fast, and having good accuracy, a YOLOv5 based solution was proposed to use deep learning for recognizing sign language [14]. In the experiments that followed, 95% precision, 97% recall, 98% mAP@0.5 and 98%@mAP@0.5:0.95 was observed which was considered adequate for real-time use of this system.

In [15], attention is brought to the COVID 19 pandemic. This paper proposed a new method based on YOLOv5 for recognizing whether people entering public places such as retail stores are wearing their face coverings correctly or not. It proposed a system where a person just needs to stand in front of a camera at the entrance and if correctly worn face covering was detected, the gate would open. An accuracy of 97.9% was achieved post testing.

Due to missing alarms and false alarms induced by onshore ship-like objects and nearby ship arrangement, nearshore ship detection is a major difficulty. A recently released research [16] suggests a Yolov5-based approach for detecting nearshore ships. The detection network incorporated the attention model and the Circle Smooth Label (CSL) to improve the accuracy. The detection experiment was then carried out using Yolov5. The Yolov5 rotation detection network was recreated using the attention model, which was integrated with the CSL algorithm to improve the network. The detection network's test result for inshore targets had mAP above 80%, confirming the capability of the CSL+Yolov5 algorithm for rotation detection.

III. DATASET DESIGN

The dataset used in the project is sourced from real world consumer sales data and consists of 1000 images of receipts from multiple retailers. This dataset contains one training set (D_{train}) and 6 testing sets(D0-D5). The first test set (D_{test}) consists of regular images with varying BRISQUE [5] scores. Test sets D1-D5 consists of images with a BRISQUE score within a specific range. The dataset description in Table I

consists of the dataset name, the description, and the number of images in the dataset while the classes to be detected are detailed in Table II.

TABLE I. DATASET DESCRIPTION

Dataset	Description	Size
D_{train}	Training Dataset	700 images
D_{val}	Validation Dataset	150 images
D_{test}	Test Dataset	150 images
D1	BRISQUE 0-20	30 Images
D2	BRISQUE 21-40	30 Images
D3	BRISQUE 41-60	30 Images
D4	BRISQUE 61-80	30 Images
D5	BRISQUE > 80	30 Images
D_{R1}	Retailer 1 Training Dataset	700 images
D_{R2}	Retailer 2 Training Dataset	700 images
D_{R1test}	Test Dataset for Retailer 1	150 images
D_{R2test}	Test Dataset for Retailer 2	150 images

TABLE II. CLASS DESCRIPTION

No	Class	Description
1	Logo	Represents the store or institution logo on top of the receipt
2	Products	Represents the table of products that appears on the receipt
3	Total	Represents the Payment information

A. BRISQUE

The Blind/Reference-less image spatial quality evaluator (BRISQUE) [5] is a spatial domain algorithm. The algorithm solely evaluates the image's 'naturalness' (or lack thereof) owing to the existence of aberrations. BRISQUE does not calculate distortion-specific characteristics such as ringing, blur, or blocking, but instead employs scene statistics of regionally normalized luminance coefficients to quantify potential losses of "naturalness" in the image due to the presence of distortions, resulting in a holistic measure of quality.

B. Image Annotation and Classification

Three classes of objects (logo, products, payment) are to be detected in an image. The ground truth of the objects will be annotated in the training images. An opensource program named LabelImg is use for the annotation of the images in the datasets. Table II consists of the description of the classes.

C. Evaluation Strategy

The evaluation technique comprises matching the anticipated class label for the ground truth bounding box and ensuring that the predicted bounding box has more than 50% Intersection over Union (IoU) area. Precision and recall are both calculated by comparing Intersection over Union (IoU), which is defined as the ratio of the area overlap between predicted and ground-truth bounding boxes divided by the area of their union. The Mean F1 Score measure is used to evaluate the match. The F1 score, calculated using (1), which is extensively employed in data analysis, assesses accuracy by combining the precision(p) and recall(r) statistics.

$$F1\ score = 2 \times \frac{p \times r}{p + r} \tag{1}$$

Precision is defined as the ratio of true positives (TP) to expected positives (TP + FP) and is calculated using the formula in (2).

$$Precision(p) = \frac{TP}{TP + FP} \tag{2}$$

Recall is defined as the ratio of true positives to actual positives (TP + FN) and is calculated using the formula in (3).

$$Recall(r) = \frac{TP}{TP + FN} \tag{3}$$

Increasing the F1-score assures a reasonable level of accuracy and recall. To get a detailed view of the results, the average precision (AP) for each of the classes is also defined and used to compare the results.

IV. METHODOLOGY

A. Model

It is critical that object detection occurs fast and accurately. Object-detection algorithms are constantly being improved by researchers. Many methods have been created and are still being developed in this area. YOLOv5 [1] is a proven algorithm in terms of speed and accuracy from the You Only Look Once (YOLO) series of CNN-based algorithms. Later, additional versions were created to increase the criteria for accuracy and speed. The final version is YOLOv5. Because of its effectiveness in detecting items, YOLOv5 was employed in this study. The architecture of the YOLOv5 algorithm can be seen in Figure 1.

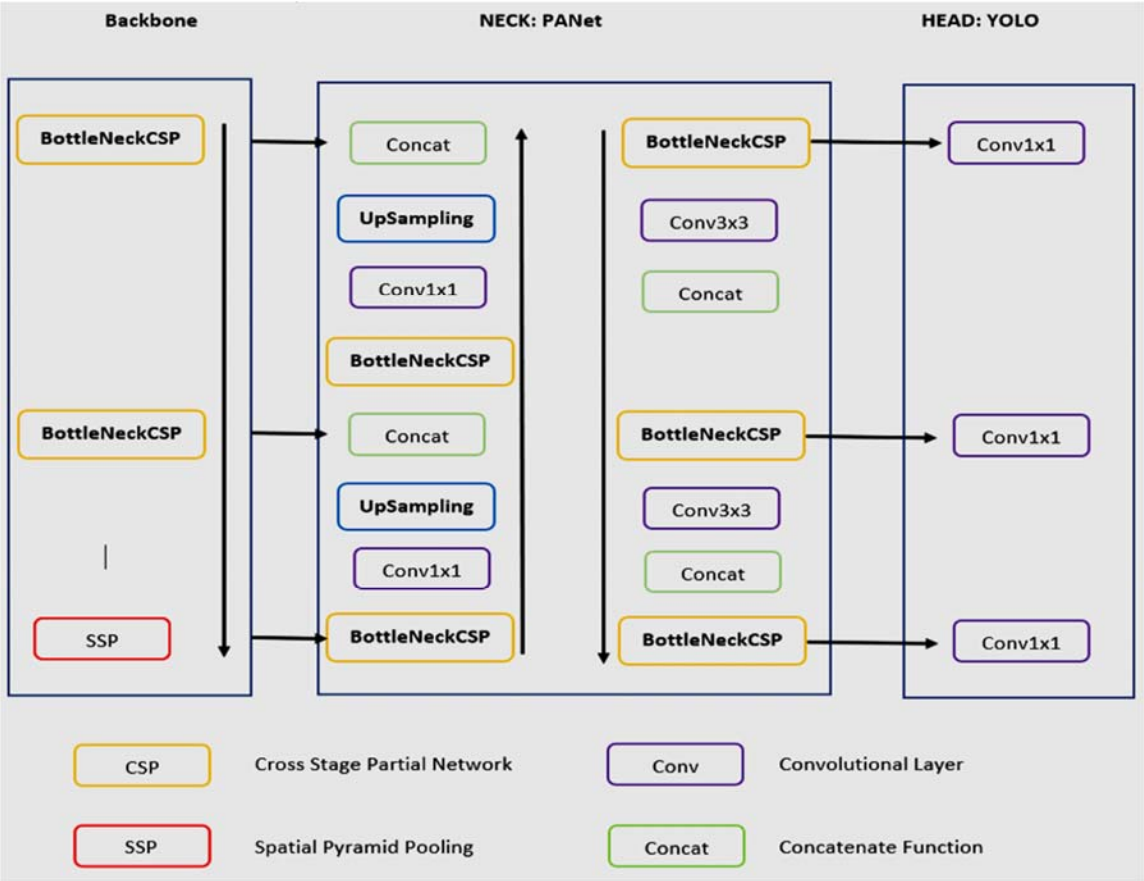

Figure 1. YOLOv5 architecture

YOLOv5's pipeline is divided into three stages. The initial step is to extract features. The CSP network is utilized in the backbone section. CSPNet [17] handles recurring gradient information issues on large-scale backbones and incorporates gradient changes into the feature map, decreasing model parameters. This guarantees that the extraction is quick and accurate.

In the second phase, the information flow in YOLOv5 is improved by making use of the path aggregation network (PANet) [18]. A new bottom-up approach that increases low level feature diffusion has been employed by the PANet. Important information at each attribute level is transported immediately to the subnet below, and this instantaneous transfer is guaranteed by the use of the adaptive feature pool, which connects the feature grid and all attribute levels.

The predictions are performed in the third phase of the model. The head used in YOLOv3 [19] is also utilized here. Feature maps of three different sizes are generated for multi scale predicting. By doing this, the model is able to detect small, medium, and even large size objects.

The YOLOv5s model was used in this project, primarily because of the speed that it offers. The high-level data flow is shown in figure 2.

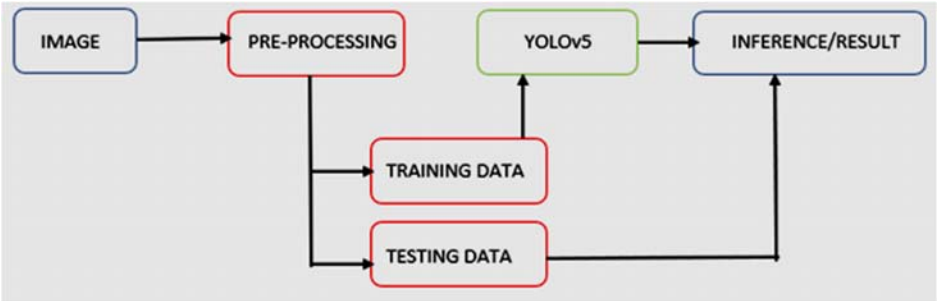

Figure 2. High Level Flow Diagram

B. Experiments

B1. Data Pre-Processing: Most ML models' success is determined on the quantity and variety of data. Data augmentation is a method of creating data for machine learning (ML) models. Data augmentation is effective for improving the performance and results of machine learning models by generating fresh and diverse instances for training datasets. When a machine learning model's dataset is rich and sufficient, the model performs better and is more accurate. Data collection and labeling may be time-consuming and expensive operations for machine learning models. Transformations in datasets employing data augmentation techniques enable for cost savings. The data augmentation steps like resize, brightness, orientation, noise was used to enhance the dataset. This resulted in a rich and diverse dataset for training.

The second stage of pre-processing was to prepare test sets to observe the effects of image quality on the performance of the model. BRISQUE [5] score was used to parameterize the quality of an image. Noise and Blur were artificially introduced in the images to achieve a higher BRISQUE score. Five test set were made consisting of images with the BRISQUE score in the range on 0-20, 21-40, 41-60, 61-80, and all images with scores above 80, respectively.

B2. Model per Retailer: YOLOv5 models was trained specifically to fit the data from each of the different receipt types in the dataset. The model's accuracy was expected to improve with two distinct models dedicated to receipts from Retailer 1 and Retailer 2. This experiment was conducted to evaluate the feasibility of creating a specialized model for use by a single organization instead of a generalized model.

B3. Geneeralized Model: A YOLOv5 model was trained on the entire dataset including different types of receipts in an attempt to create a generalized model and the results were compared to those of the individual models created for Retailer 1 and Retailer 2.

B4. Image Quality: A YOLOv5 model trained on the entire dataset was used to get inferences on the five datasets with varying image quality to observe the effects of the image quality on the performance of the model.

B5. Variants of YOLOv5: Three variants of the YOLOv5 algorithm, YOLOv5s, YOLOv5m and YOLOv5l were trained using the same training dataset and inferences were obtained on the test set to observe the improvements in performance if any.

V. RESULTS

A. Model Per Receipt Type

A Yolov5s model was trained on data from 2 different retail stores, R1 and R2, respectively. For R1, dataset D_{R1} was used to train the model while for R2, D_{R2} was used to train the model. Datasets D_{R1test} and D_{R2test} were used to test the performance of each model respectively. The generalized model was trained using the training set D_{train}. The observed results are documented in Table III.

TABLE III. SPECIALIZED AND GENERAL MODEL RESULTS

Receipt Type	Precision	Recall	F1 Score	mAP@.5
R1	0.899	0.948	0.923	0.941
R2	0.872	0.767	0.816	0.795
Generalized Model	0.908	0.840	0.873	0.866

The models for Retailer 1(R1) and Retailer 2(R2) were each trained for 350 epochs. As seen in Table III, the model specific to R1 performed the best while the generalized model performed the worst. This shows that for solutions intended for a single entity, a specific model will have superior performance as compared to a generalized model.

B. Image Quality

The BRISQUE [5] score is used to parameterize the quality of an image. The generalized model trained on dataset D_{train} was used. The observed difference in performance when running inference on different datasets is documented in Table IV.

TABLE IV. IMAGE QUALITY RESULTS

Dataset	Description	Precision	Recall	F1 Score
Baseline	All training images	0.899	0.948	0.923
D1	BRISQUE 0-20	0.936	0.972	0.954
D2	BRISQUE 21-40	0.922	0.949	0.935
D3	BRISQUE 41-60	0.811	0.721	0.763
D4	BRISQUE 61-80	0.576	0.538	0.556
D5	BRISQUE > 80	0.486	0.585	0.531

As seen from the graph in Figure 2, the performance of the model is significantly reduced for the datasets D3, D4 and D5. This shows that for images with BRISQUE score above 40, the accuracy of the model is greatly reduced. It is also observed that for datasets D1 and D2, the performance

showed improvement of 3.1% and 1.2% respectively. While looking at the F1 scores for each class over the different test sets, it is observed that the most significant decrease in performance is taken by the Total class, whereas the Logo and Products class while losing performance, have relatively good performance compared to the Total class. The results detailed in Table V and depicted in Figure 3.

TABLE V. INDIVIDUAL CLASS RESULTS

Dataset	Logo	Products	Total
D0	0.965	0.843	0.955
D1	1	0.897	0.963
D2	0.846	0.979	0.981
D3	0.755	0.816	0.67
D4	0.963	0.695	0.547
D5	0.794	0.776	0.12

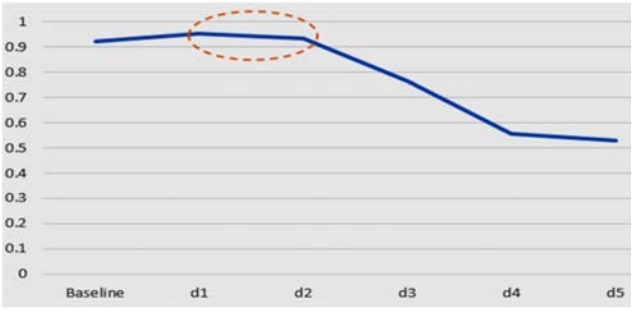

Figure 3. F1 score across datasets D1 - D5

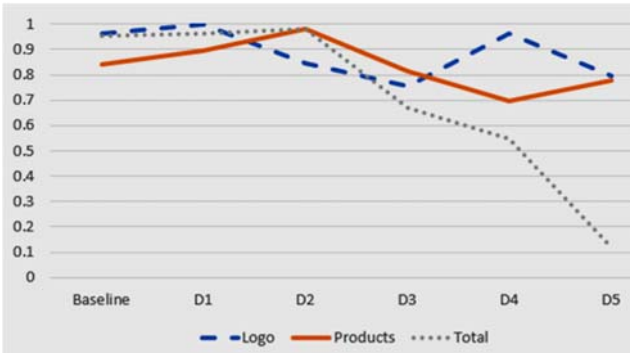

Figure 4. F1 score (per class) across datasets D1 - D5

C. Variants of YOLOv5

Three variants of the YOLOv5 algorithm, YOLOv5s (small), YOLOv5m (medium) YOLOv5l (large) were trained on dataset $D_{train.}$ The results observed are detailed in Table VI.

TABLE VI. VARIANTS OF YOLOV5

Model	Precision	Recall	F1
YOLOv5s	0.899	0.948	0.923
YOLOv5m	0.948	0.964	0.956
YOLOv5l	0.941	0.892	0.916

The YOLOv5s model, which has the least number of trainable parameters and layers, achieved a f1 score of 0.923 and mAP of 0.941. The YOLOv5m achieved a f1 score of 0.956 and a mAP of 0.970 while the YOLOv5l achieved a f1 score of 0.916 and a mAP of 0.935. Of the three model, the YOLOv5m provided the highest performance while the YOLOv5l provided the least performance.

VI. CONCLUSION AND FUTURE WORK

This study detects three objects, Logo on the receipt, the products table listed on the receipt and the payment details for the receipt, respectively. For this task, the latest algorithm in the YOLO series, which is the YOLOv5 algorithm, was used. The effects of degrading image quality on the performance of the YOLOv5 algorithm were also quantified by making use of the BRISQUE score. It was observed that for datasets D3, D4 and D5, there was a significant drop in performance, indicating that while running inferences on images with BRISQUE score above 40, the accuracy of the model starts to drop.

This paper utilizes the real-world invoice images provided by the NPD Group. Structuring the data after running an OCR engine on an image is a big challenge. The output of an OCR engine is often an editable text file. While advancements are being made in OCR technology, these text files often need proof reading and proper formatting to get the desired result. Future work will be focused on implementing Optical Character Recognition (OCR) technology in tandem with the Object Detection algorithm to streamline the data extraction process of invoices. The object detection algorithm will be implemented to correctly identify and draw bounding boxes around textual data required in the input image. The OCR engine will then be implemented to scan characters only inside the bounding boxes provided by the object detection algorithm. The aim is to identify and implement the ideal combination of an object detection algorithm and an OCR engine for data extraction from invoices. The results observed in this paper show that an object detection algorithm like YOLOv5 can be successfully used in identifying textual objects from an input image while using a relatively small dataset and a short training period. Using a large and diverse dataset and training the algorithm for longer periods will result in a generalized model suitable for most invoice types and formats.

The use of BRISQUE score in this paper as a parameter for image quality raises the possibility of a system that dynamically chooses a model for object detection and discards images that have distinctively poor quality before the OCR engine is run. Investigation into this possibility will also take place in the hopes of reducing the processing power required and making the entire system more efficient.

REFERENCES

[1] *ultralytics. yolov5*, June 2020, [online] Available: https://github.com/ultralytics/yolov5.

[2] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, et al., "Ssd: Single shot multibox detector", *European conference on computer vision*, pp. 21-37, October 2016.

[3] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation", *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580-587, 2014

[4] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation", *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580-587, 2014

[5] A. Mittal, A. K. Moorthy and A. C. Bovik, "No-Reference Image Quality Assessment in the Spatial Domain," in IEEE Transactions on Image Processing, vol. 21, no. 12, pp. 4695-4708, Dec. 2012.doi: 10.1109/TIP.2012.2214050

[6] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, et al., "Microsoft COCO: Common objects in context", *European conference on computer vision*, pp. 740-755, 2014.

[7] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, et al., "Deep learning for generic object detection: A survey", *International journal of computer vision*, vol. 128, no. 2, pp. 261-318, 2020.

[8] J. Redmon and A. Farhadi, "YOLO9000: Better faster stronger", *CVPR*, 2017

[9] T.-Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollar, "Focal loss for dense object detection", *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2380-7504, 22–29 October 2017.

[10] K. He, G. Gkioxari, P. Dollar and R. Girshick, "Mask r-cnn", *IEEE International Conference on Computer Vision (ICCV)*, pp. 2980-2988, 2017.

[11] T.-Y. Lin, P. Dollar, R. B. Girshick, K. He, B. Hariharan and S. J. Belongie, "Feature pyramid networks for object detection", *CVPR*, vol. 1, pp. 4, 2017.

[12] Margrit Kasper-Eulaers et al., "Detecting Heavy Goods Vehicles in Rest Areas in Winter Conditions Using YOLOv5", *Algorithms*, vol. 14.4, pp. 114, 2021

[13] J. Tang, S. Liu, B. Zheng, J. Zhang, B. Wang and M. Yang, "Smoking Behavior Detection Based On Improved YOLOv5s Algorithm," *2021 9th International Symposium on Next Generation Electronics (ISNE)*, 2021, pp. 1-4, doi: 10.1109/ISNE48910.2021.9493637.

[14] T. F. Dima and M. E. Ahmed, "Using YOLOv5 Algorithm to Detect and Recognize American Sign Language," *2021 International Conference on Information Technology (ICIT)*, 2021, pp. 603-607, doi: 10.1109/ICIT52682.2021.9491672.

[15] G. Yang *et al*., "Face Mask Recognition System with YOLOV5 Based on Image Recognition," *2020 IEEE 6th International Conference on Computer and Communications (ICCC)*, 2020, pp. 1398-1404, doi: 10.1109/ICCC51575.2020.9345042.

[16] Q. Fu, J. Chen, W. Yang and S. Zheng, "Nearshore Ship Detection on SAR Image Based on Yolov5," *2021 2nd China International SAR Symposium (CISS)*, 2021, pp. 1-4, doi: 10.23919/CISS51089.2021.9652233.

[17] C. Y. Wang, H. Y. M. Liao, Y. H. Wu, P. Y. Chen, J. W. Hsieh and I. H. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN", *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 390-391, 202.

[18] K. Wang, J. H. Liew, Y. Zou, D. Zhou and J. Feng, "Panet: Few-shot image semantic segmentation with prototype alignment", *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9197-9206, 2019.

[19] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement", *arXiv preprint*, 2018.

.