Performance Analysis of Object Detection Algorithms on YouTube Video Object Dataset

Chethan Sharma, Siddharth Singh, Poornalatha G*, Ajitha Shenoy KB

Abstract— **Object Recognition is a terminology used to refer to a collection of computer vision tasks that are involved in object identification in digital images and videos. In this paper, different object detection algorithms were implemented on Youtube object dataset. Each object detection algorithm has its own advantages and limitations which depend on the dataset used. It was observed that YOLO and SSD, being state-of-art algorithms, demonstrate better performance than other models on youtube video object dataset. SSD is better at detecting smaller objects. Centernet performs poorly on this dataset.**

***Index Terms*— *Average Precision*, CenterNet, DETR, Object Recognition, SSD, YOLO.**

I. Introduction

Object Recognition is a terminology used to refer to a collection of computer vision tasks that are involved in object identification in digital images and videos. Object recognition includes tasks of both image classification and object detection. Image classification is a task of determining class of the objects present in the image and object detection means localizing the objects in an image. Localization here refers to finding the location of the objects in the image or video by drawing the bounding box around the objects. One extension to these tasks of computer vision is object segmentation where instead of bounding box, all the pixels of the objects are highlighted. In most of the computer vision systems, first task that is performed is object detection. Once the objects in a particular scene or an image is detected, it is possible to obtain more information about the object such as specific instance of the object, tracking of the object over a sequence and extracting further information about the object and also infer the presence of other objects in the same image and also other contextual information about the scene or the image. Applications of object detection are wide in range which includes retrieval, robotics, consumer electronics etc. Even though the field of object detection has seen improvements over the years, it is still a open research area in computer vision. Each technique has its own advantages and disadvantages and the analyst can generally pick the strategy that suits their necessities best.

II. Literature Review

In [1], YOLO, faster-RCNN and fast-RCNN were used for people counting. Only YOLO is used and it is selected over the other algorithm because of low computation overhead. YOLO-PC outperforms YOLO as it re-trains and is capable of ignoring irrelevant objects. In [2], a YOLO based algorithm was designed for the purpose of pedestrian detection, with a new network structure called the Three Passthrough layers. This innovation solves the disappearing gradient problem causing inaccurate detections. It was also discovered that YOLOv2 struggles with small object detection due to the loss of the finer features of the input image due to downsampling. This issue, however, being resolved in the case of YOLOv3, the authors claim that YOLOv3 can improve the results further. In [3], YOLO was used to detect bones in the pelvic area, and it achieved better accuracy than fast-RCNN. The authors further claim that the results can be improved with the use of SSD or even YOLOv3. In [4], the authors tried real-time face detection using YOLO. YOLO was selected over SSD and faster-RCNN because of its faster processing speed, with nearly the same accuracy as the other two. YOLOv3 succeeds where YOLOv2 fails, i.e, in detecting small objects. But, as the authors also found out, RetinaNet still surpasses YOLOv3 in that regard.

For the purposes of fast video querying, the authors [5] proposed a new lightweight object detector FDet, which achieves 29.7 AP on COCO benchmark with lighter backbone as compared to YOLO or SSD. Anchor-free detectors are better than anchor-based detectors as they avoid complicated computation and hyper-parameters related to anchor boxes. However, this theory was invalidated by the authors as it was found that CenterNet (which is anchor-free) performed worse than other anchor-based detectors. In [6], a fault detection technique for catenary systems using modified CenterNet achieved a 94 accuracy on their dataset with ResNet-101 backbone. Their experiments show that CenterNet with a light backbone such as ResNet-34 can detect small targets with good accuracy and detection rate. Since current carrying rings are usually small targets, models like RetineNet can be used. In [7], CenterNet was used for the detection of intersection of roads. However, CenterNet was selected over YOLO, SSD and Faster-RCNN because of its high detection precision for

Manuscript received January 06, 2021; revised February 15, 2021.

Chethan Sharma is working as an Assistant Professor- Senior in Department of Information and Communication Technology, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka, India. (e-mail: chethan.sharma@manipal.edu).

Siddharth Singh is a final year student of the Department of Information and Communication Technology, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka, India. (e-mail: siddharthsingh1311@gmail.com)

Poornalatha.G is working as an Associate Professor-Senior in the Department of Information and Communication Technology, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka, India. (Corresponding Author, e-mail: poornalatha.g@manipal.edu).

K.B. Ajitha Shenoy is working as an Associate Professor-Senior in the Department of Information and Communication Technology, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka, India. (e-mail: ajith.shenoy@manipal.edu)

small targets. It achieved an accuracy of 96.2. Also it was shown that CenerNet takes a lot less time for training than other models (namely, YOLOv3, Faster-RCNN). Using the TIMIT corpus dataset, Center was used in [8], for phoneme recognition using transfer learning. It achieved 15.89 error rate. CenterNet achieves better than YOLOv3 in speech recognition tasks, as discovered by the authors. In [9], a new framework, Few-Shot Detector, FSD, based on meta learning, to avoid large training time needed for object detection. The proposed framework involves a meta-learner and an object detector. SSD was selected because of its simplicity and effective multiscale detection with auxiliary convolution structure.

In [10], Mask-SSD and Mask-RCNN were used for instance segmentation and achieved almost similar performances. However, Mask-SSD's inference time is less. It achieved an AP of 39.3 on the dataset. The authors demonstrated that single stage detection techniques can achieve the same accuracy as two stage techniques and have relatively low inference time. In [11], SSD was modified for the task of licence plate recognition. A lighter, custom backbone, which requires less parameters than VGG, was used for boosting the inference speed of the model. It achieved 99.79 segmentation and recognition accuracy on UCSD stills dataset. However, in [12], a new backbone for SSD was proposed, called DensenNet. DenseNet replaces the original feature extractor VGG16 and achieves a mAP of 29.5 on the COCO dataset. The model requires exactly ½ and 1/9 of the total parameters as compared to SSD and Faster-RCNN. The author also introduced a feature-fusion module which improves the detection accuracy on small targets.

III. Methodology and model architectures

In this section we brief about the architectures of different models and algorithms used for performance analysis on the YouTube video object dataset.

A. *YOLO V3*

YOLOv3 has a 106 layer fully convolutional architecture which performs multilabel classification for objects detected in images. YOLOv3uses a variant of Darknet . The most prominent feature of v3 is that the detections are made at three different scales. 1x1 kernel is applied on the feature maps to perform object detection. The object detection is done by applying 1 x 1 detection kernels on feature maps. The shape of the detection kernel is 1 x 1 x (B x (5 + C)), where B represents the number of bounding boxes predicted by a cell of the feature map, C represents the number of classes and 5 represents a object confidence and attributes of bounding boxes which are 4 in number. In YOLOv3trained on COCO, B = 3 and C = 80, so the kernel size is 1 x 1 x 255. For our study we have used YOLOv3 pretrained on MS COCO dataset, with a backbone of darknet-53[13] as a feature extractor, in pre-processing step first the image is rescaled. In this step the shorter side is converted 512px keeping the aspect ratio intact. The input to the model is in the form of tensor arrays[14], which are obtained by normalizing the rescaled images.

B. *SSD*

Single Shot Multibox Detector is a one stage object detection algorithm i.e. it does not have any proposal generation. It is based on feed forward neural network and produces a fixed-size collection of bounding boxes with scores for each class. Predicting the category scores and box offsets for bounding boxes are the core of SSD. To achieve high detection accuracy prediction, feature maps of different scales are produced and separated by aspect ratio. Initially a standard neural network, used for images classification is used, truncated before the classification layer, this network is called the base network, then a neural network to produce detection is used with the default feature.

In [15], SSD is implemented with VGG16 as the base network. But for our experiment we have used SSD with ResNet-50 as the base network.

C. *DETR*

Detection Transformer (DETR) [16] proposes a new Transformer based method for solving object-detection tasks. DETR views object detection as a direct set prediction problem. It is the first use of transformers as a central building block in object detection problems. Transformers have been used very effectively in sequential data problems such as NLP. Transformers rely on a simple mechanism known as attention. DETR performs a global reasoning on Image with help of the self-attention mechanism of transformers. DETR achieves almost the same accuracy as Faster-RCNN with a relatively simple pipeline.DETR employs a very simple architecture as compared to the current state-of-the-art models. It has three main components first is a CNN based feature extractor, then a Transformer encode-decoder and then a simple Feed Forward Network for making the final predictions. As explained in [17], like other object detectors DETR uses a CNN backbone for generation of a low-resolution input map. Then the spatial dimensions of the input map are collapsed into 1-dimension as the encoder takes sequential data. Each encoder consists of a self-attention module and Feed Forward Network encoders are input invariant. Decoder in the transformer decodes a N of input embeddings at each decoder layer. The input embeddings in the decoder are permutation invariant and each produce a different result. These input embeddings are referred to as object queries. These object queries are decoded into output embedding and that are then converted into class labels and bounding boxes. At the final stage a Feed Forward Network is used to make the final predictions.

D. *CenterNet*

CenterNet [17] is the successor of CornerNet [18]. The CornerNet, overcomes the limitations associated with anchor-boxes which are used frequently in many state-of-the-art single-shot detectors. Anchor boxes are boxes of various sizes and aspect ratios and possess two major drawbacks: A large set of anchor boxes is required typically, and secondly, design becomes complicated because of the introduction of various hyperparameters. CornerNet uses a pair of corner-key points to represent the bounding box (the top-left and bottom-right corners) of the object detected thereby simplifying the design. CenterNet improves over CornerNet by exploring the central part of a proposal, which is the region that is close to the geometric center of the object. This leads to another center keypoint in addition to the corner keypoints..The center keypoint helps the network by not misgrouping pairs of keypoints, and thus not generating extra bounding boxes overlapping with the boundary of the object but not containing it.

The fundamental concept is that objects are treated as center points and thus eliminate any bounding boxes that do not contain these points. By forcing the model to detect the center point, the model ensures faster performance and improved accuracy as CenterNet simply extracts the center point per object without the need for post-processing or grouping in contrast to CornerNet. Furthermore, the original corner pooling was enhanced to cascade corner pooling which allows the algorithm to perceive internal information, by obtaining the max summed response in both the boundary and internal directions of objects on a feature map for predicting corners. For prediction of center keypoints, center pooling is used to calculate the max summed response in both the horizontal and vertical direction of center keypoint on a feature map.

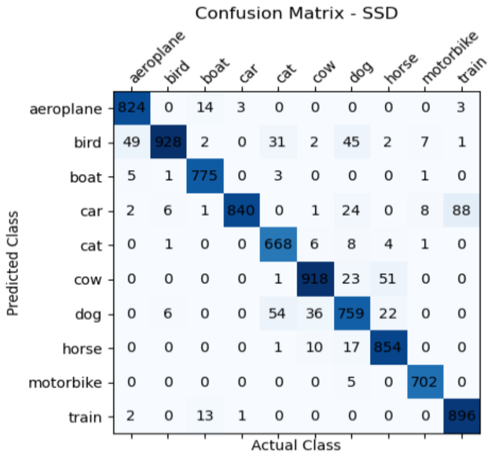
IV. Results and Discussion

A. Dataset

The dataset used in our study is a subset YouTube Object Detection [20][21] dataset. The dataset is created from YouTube videos of 10 object classes of PASCAL VOC Challenge. Approximately, 1000 images for each of the 10 classes were randomly sampled from the total of 720,000 frames.

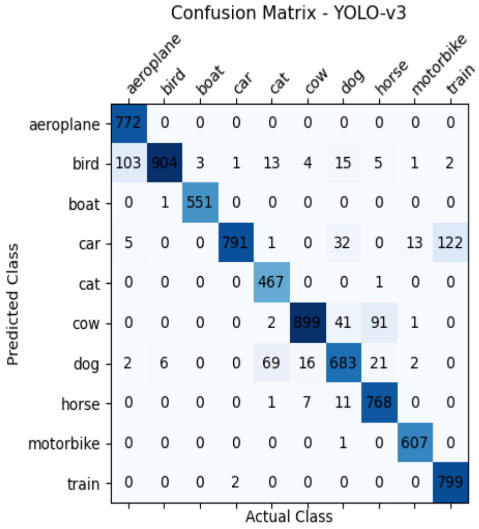
B. Confusion Matrix

A Confusion Matrix is a table used for describing the performance of classification model. In Multi-Class Confusion Matrix, the diagonal elements represent all the correct classification. From the Confusion Matrix the precision for any class say C, can be calculated as the ratio of the number of correct classification that is, value in the matrix at the intersection row C and column C to all the samples that are classified as class C, that can be calculated as the sum of all the values in row C. And recall for any class can be calculated as the ratio of number of correct classifications to the total number samples which are from class C which is the sum of all the values in the column C. Figures 1 to 4 shows the confusion matrixes obtained from our experiments.



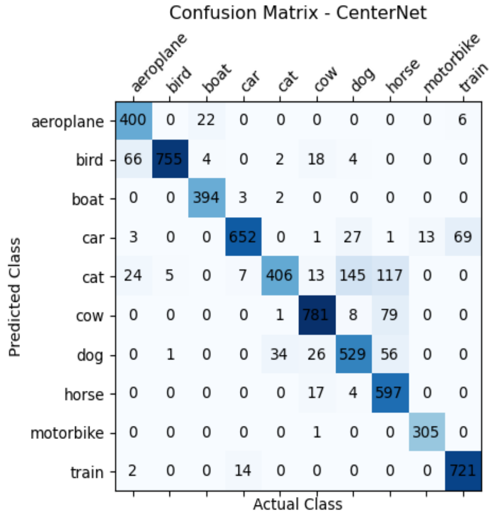
	aeroplane	bird	boat	car	cat	cow	dog	horse	motorbike	train
aeroplane	824	0	14	3	0	0	0	0	0	3
bird	49	928	2	0	31	2	45	2	7	1
boat	5	1	775	0	3	0	0	0	1	0
car	2	6	1	840	0	1	24	0	8	88
cat	0	1	0	0	668	6	8	4	1	0
cow	0	0	0	0	1	918	23	51	0	0
dog	0	6	0	0	54	36	759	22	0	0
horse	0	0	0	0	1	10	17	854	0	0
motorbike	0	0	0	0	0	0	5	0	702	0
train	2	0	13	1	0	0	0	0	0	896

Figure 1: Confusion Matrix – SSD



	aeroplane	bird	boat	car	cat	cow	dog	horse	motorbike	train
aeroplane	772	0	0	0	0	0	0	0	0	0
bird	103	904	3	1	13	4	15	5	1	2
boat	0	1	551	0	0	0	0	0	0	0
car	5	0	0	791	1	0	32	0	13	122
cat	0	0	0	0	467	0	0	1	0	0
cow	0	0	0	0	2	899	41	91	1	0
dog	2	6	0	0	69	16	683	21	2	0
horse	0	0	0	0	1	7	11	768	0	0
motorbike	0	0	0	0	0	0	1	0	607	0
train	0	0	0	2	0	0	0	0	0	799

Figure 2: Confusion Matrix – YOLO V3



	aeroplane	bird	boat	car	cat	cow	dog	horse	motorbike	train
aeroplane	400	0	22	0	0	0	0	0	0	6
bird	66	755	4	0	2	18	4	0	0	0
boat	0	0	394	3	2	0	0	0	0	0
car	3	0	0	652	0	1	27	1	13	69
cat	24	5	0	7	406	13	145	117	0	0
cow	0	0	0	0	1	781	8	79	0	0
dog	0	1	0	0	34	26	529	56	0	0
horse	0	0	0	0	0	17	4	597	0	0
motorbike	0	0	0	0	0	1	0	0	305	0
train	2	0	0	14	0	0	0	0	0	721

Figure 3: Confusion Matrix – CenterNet

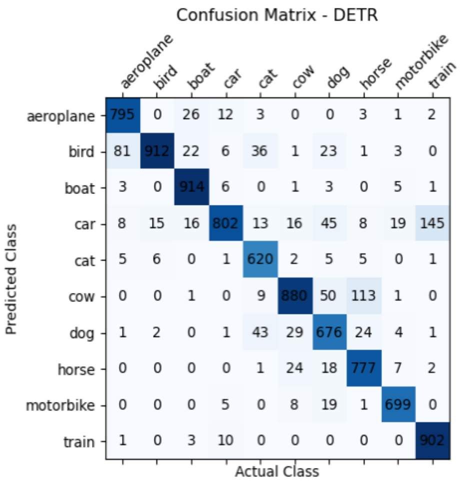


Figure 4: Confusion Matrix - DETR

C. *Mean Average Precision*

Location of the object is defined by the bounding box. For evaluation of detection performance, we have used PASCAL VOC Challenge metric, we first define Intersection over Union (IoU), Intersection over Union is a metric that evaluates the overlap between two bounding boxes. Suppose BGround-Truth be the original bounding boxes generated by humans and BPredicted is the bounding boxes predicted by the model. Intersection over union is defined as:

$$IoU = \frac{area(B_{Ground-Truth} \cap B_{Predicted})}{area(B_{Ground-Truth} \cup B_{Predicted})}$$

Then, we define following cases for IOU :

i) If IoU > 0.5 , Detection will be classified as True Positive.

ii) If IoU < 0.5, Detection will be classified as False Positive.

iii) If the detector fails to recognize the object present in the image. It is classified as False Negative.

These three cases are then used in the calculation of Precision and Recall. Precision represents the percentage of correct predictions and Recall represents the percentage of True Positives among all the ground-truth labels.

Using these metrics precision-recall curve for every class is plotted and the area under each curve is calculated, as the curve is piecewise-constant, and no approximations are needed. Area under each curve represents the values of Average Precision (AP) for that class. After mean of the all the AP values is calculated and this is the final mAP for that model, which will be used for the performance comparison. Figures 5 to 8 shows Average-Precision graphs of our experiment.

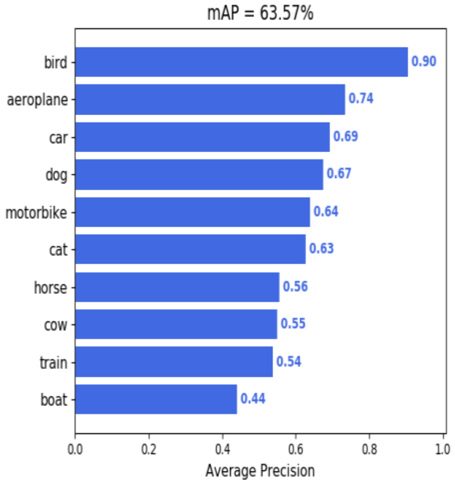


Figure 5: Average-Precision for- SSD

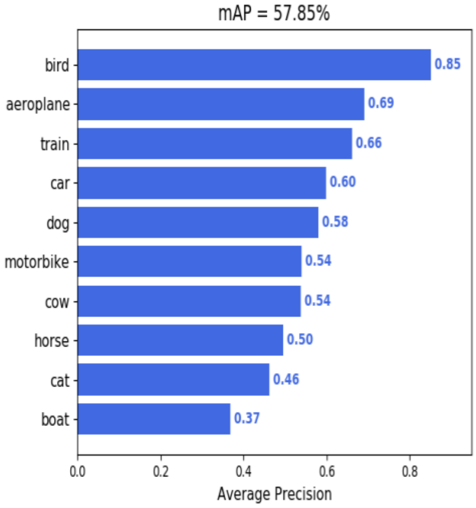


Figure 6: Average-Precision for- YOLO

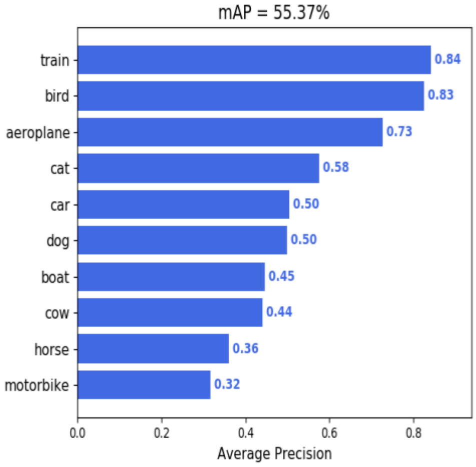


Figure 7: Average-Precision for- DETR

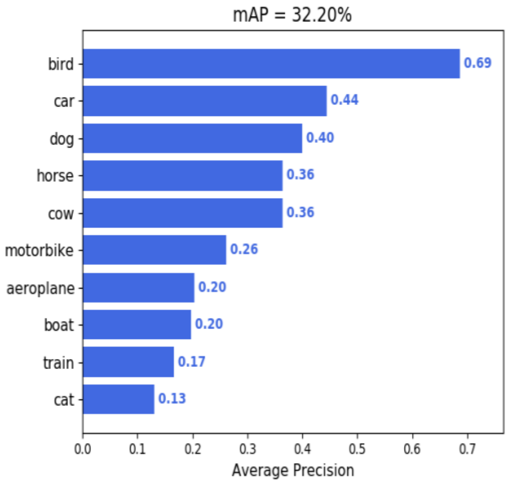


Figure 8: Average-Precision for- CenterNet

Average Precision curves show that SSD has achieved best mAP on the dataset, followed by YOLOv3 and DETR with almost comparable performance, while CenterNet has achieved the lowest mAP. The bird class has high AP value for every model. This may be due to two factors, first relatively low variance for that class in the dataset taken. And secondly, because of higher quality frames compared to other classes. It is also observed that CenterNet has performed poorly on the images in which a single object covered a large area of the image or the images in which the object was not completely present in the frame. On the other hand, DETR has performed quite well on such images. On observing the average precision values for every model, we see that SSD and YOLO-v3 have performed well on every class and do not show any bias towards any class. While CenterNet has very good AP value for one class, others are relatively low. Finally the results are summarized in table 1.

Table 1. Mean Avergae Precision of Models

Model	mAP
CenterNet	32.20 %
SSD	63.57 %
YOLOV3	57.85 %
DETR	55.37 %

V. Conclusion

YOLO and SSD, being state-of-art algorithms, demonstrate better performance than other models. SSD is actually better at detecting smaller objects. DETR, though being slightly inferior to SSD and YOLO, outperforms CenterNet by a long stride. However, it must be taken into account the limitations of CenterNet, which makes it unable to classify properly. It was also observed that some classes were easily recognized, and some were not. It can be concluded that the dataset presented variance in the classes on which the models were not previously trained, and some classes had the same distribution as the models had been trained to detect. Another finding of this study, at least for the dataset used in this study, is that some models tend to be better at detecting some classes and not others. Further study is required to say whether each and every model has a certain bias towards some classes.

References

[1] P. Ren, W. Fang and S. Djahel, ”A novel YOLO-Based real-time people counting approach,” 2017 International Smart Cities Conference (ISC2), Wuxi, 2017.

[2] W. Lan, J. Dang, Y.Wang and S.Wang, ”Pedestrian Detection Based on YOLO Network Model,” 2018 IEEE International Conference on Mechatronics and Automation (ICMA), Changchun, 2018.

[3] Z. Krawczyk and J. Starzy´nski, ”Bones detection in the pelvic area on the basis of YOLO neural network,” 19th International Conference Computational Problems of Electrical Engineering, Banska Stiavnica, 2018.

[4] W. Yang and Z. Jiachun, ”Real-time face detection based on YOLO,” 2018 1st IEEE International Conference on Knowledge Innovation and Invention (ICKII), Jeju, 2018.

[5] Jiansheng Dong, Jingling Yuan, Lin Li, Xian Zhong, and Weiru Liu. “Optimizing Queries over Video via Lightweight Keypoint-based Object Detection” In Proceedings of the 2020 International Conference on Multimedia Retrieval (ICMR ’20). ACM 2020

[6] Chenchen Huang and Yuan Zeng “The fault diagnosis of catenary system based on the deep learning method in the railway industry” In Proceedings of the 5th International Conference on Multimedia and Image Processing (ICMIP ’20). ACM 2020.

[7] G. Lai, Y. Zhang, X. Tong and Y. Wu, ”Method for the Automatic Generation and Application of Landmark Control Point Library,” in IEEE Access, vol. 8, pp. 112203-112219, 2020.

[8] M. Algabri, H. Mathkour, M. A. Bencherif, M. Alsulaiman and M. A. Mekhtiche, ”Towards Deep Object Detection Techniques for Phoneme Recognition,” in IEEE Access, vol. 8, pp. 54663-54680, 2020.

[9] K. Fu et al., ”Meta-SSD: Towards Fast Adaptation for Few-Shot Object Detection With Meta-Learning,” in IEEE Access, vol. 7, pp. 77597-77606, 2019.

[10] H. Zhang, Y. Tian, K. Wang, W. Zhang and F. Wang, ”Mask SSD: An Effective Single-Stage Approach to Object Instance Segmentation,” in IEEE Transactions on Image Processing, vol. 29, pp. 2078-2093, 2020.

[11] R. D. Castro-Zunti, J. Y´epez and S. Ko, ”License plate segmentation and recognition system using deep learning and OpenVINO,” in IET Intelligent Transport Systems, vol. 14, no. 2, pp. 119-126, 2, 2020.

[12] S. Zhai, D. Shang, S. Wang and S. Dong, ”DF-SSD: An Improved SSD Object Detection Algorithm Based on DenseNet and Feature Fusion,” in IEEE Access, vol. 8, pp. 24344-24357, 2020.

[13] Joseph Redmon. (2013–2016). Darknet: Open Source Neural Networks in C. http://pjreddie.com/darknet/.

[14] Chen, T., Li, M., Li, Y., Lin, M., Wang, N., Wang, M Zhang, Z “ Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems” arXiv preprint arXiv:1512.01274, 2015

[15] Wei Liu et.al, “SSD: Single Shot MultiBox Detector” CoRR, abs/1512.02325, 2015

[16] J Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S “End-to- End Object Detection with Transformers” , 2020.

[17] Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., Tian, Q. (2019). Centernet: Keypoint triplets for object detection. In Proceedings of the IEEE International Conference on Computer Vision (pp. 6569-6578).

[18] Law, H., Deng, J. (2018). Cornernet: Detecting objects as paired keypoints. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 734-750).

[19] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L. “Microsoft coco: Common objects in context” In European conference on computer vision (pp. 740-755). Springer, Cham, 2014.

[20] Learning Object Class Detectors from Weakly Annotated Video Alessandro Prest, Christian Leistner, Javier Civera, Cordelia Schmid, Vittorio Ferrari, In Computer Vision and Pattern Recognition (CVPR), 2012.

[21] Analysing domain shift factors between videos and images for object detection Vicky Kalogeiton, Vittorio Ferrari, Cordelia Schmid, In PAMI, 2016.

[22] Ayoosh Kathuria, “What’s new in YOLO v3?”, www.towardsdatascience.com/yolo-v3-object-detection-53fb7d3bfe6b, 2018