COMPARATIVE STUDY OF DEEP LEARNING MODELS USING OBJECT DETECTION

Rohith G*

* Centre for Computational Engineering and Networking, Amrita Viswa Vidyapeetham, Coimbatore

Abstract- The role of sensors such as cameras or LiDAR (Light Detection and Ranging) is crucial for the environmental awareness of self-driving cars. However, the data collected from these sensors are subject to distortions in extreme weather conditions such as fog, rain, and snow. This issue could lead to many safety problems while operating a self-driving vehicle. We need object detection for automated driving. It is important that object detection be accurate overall and robust to weather and environmental conditions and run in real-time. This requires image processing algorithms to inspect the contents of the images.

The purpose of this study is to compare 2 state of art deep learning models which are Region Based Convolutional Neural Networks (RCNN) and You Only Look Once (Yolo V5) algorithms based on their Mean average Precision (Map) and Intersection over union (IOU) values. In this analysis we used a Dawn Dataset which comprises a collection of 1000 images from real-traffic environments, which are divided into four sets of weather conditions: fog, snow, rain and sandstorms. Since we are focusing only non-foggy weather data, we took only the foggy images from the data as we are only concentrating on foggy images. Collecting and processing data in adverse weather conditions is often more difficult than data in good weather conditions. The dataset is annotated with object bounding boxes for autonomous driving and video surveillance scenarios. This data helps interpreting effects caused by the adverse weather conditions on the performance of vehicle detection systems.

Index Terms- object detection, deep learning, adverse weather conditions, foggy perception.

I. INTRODUCTION

The automobile industries have developed rapidly since the first demonstration in the 1980s, the vehicle navigation and intelligence system have improved. However, the increase in road vehicles raises traffic congestion, road safety, pollution, etc. Autonomous driving is a challenging task; a small error in the system can lead to fatal accidents. Visual data play an essential role in enabling advanced driver-assistance systems in autonomous vehicles. The low cost and wide availability of vision-based sensors offer great potential to detect road incidents. Additionally, emerging autonomous vehicles use various sensors and deep learning methods to detect and classify four classes (such as a vehicle, pedestrian, traffic sign, and traffic light) to improve safety by monitoring the current road environment. Object detection is a method of localizing and classifying an object in an image to understand the image entirely. It is currently one of the first fundamental tasks in vision based autonomous driving. The object detection methods make bounding boxes around the detected objects and the predicted class label and confidence score associated with each bounding box. Object detection and tracking are challenging tasks and play an essential role in many visual-based applications. At present, the deep learning field provides several methods for advancing the automation levels by improving the environment perception. The autonomous vehicles have developed from Level-0 class with no automation to Level-1 with driver assistance automation. The Level-2 class with partial automation enables the vehicle to assist in steering and acceleration functionality, and the driver controls many safety-critical actions. For Level-3 class with conditional automation, the intelligent vehicle must monitor the whole surroundings in real-time, and the driver can only take control over the vehicle when prompted by the system. However, to achieve autonomous driving in vehicles, there is a long way to reach the Level-4 class with high automation, and finally, the Level-5 class with full automation. The sensors used to detect and monitor vehicles may be identified as containing three components: the transducer, the unit for signal processing and the device for data processing. In an autonomous vehicle, all types of sensors are essential to obtain correct information about its surrounding environment. At present sensors used in the vehicles primarily include the Monocular Camera, Binocular camera, Light detecting and ranging(LiDAR), Global Navigation Satellite System (GNSS), Global Positioning System(GPS), Radio detection and Ranging(Radar), Ultrasonic sensor, odometer and many more. However, all these sensors have benefits and drawbacks. A LiDAR operates with a similar principle of radar, but it emits infrared light waves instead of radio waves. It has much higher accuracy than radar under 200 meters. Weather conditions such as fog or snow have a negative impact on the performance of LiDAR. Another aspect is the sensor size: smaller sensors are preferred on the vehicle because of limited space and aerodynamic restraints, and LiDAR is generally larger than radar, stereo camera , flash camera , event camera , and thermal camera . However, researchers work on reducing the cost, size, and weight of

LiDAR recently , but it still needs more work. Furthermore, the radar only detects objects in close range. Therefore, the radar may detect objects at less than their specified range if a vehicle moves faster. Furthermore, both binocular cameras and monocular cameras may produce worse detection results in low light. The GPS chip is a power-hungry device that drains the battery rapidly and is costly. The GPS may also have inaccuracy due to environmental interference. Sensors are mainly used to perceive the environment, including dynamic and static objects, e.g., drivable areas, buildings, pedestrian crossings, Cameras, LiDAR, Radar, and Ultrasonic sensors are the most commonly used modalities for this task. Early autonomous driving systems heavily relied on sensory data for accurate environment perception. Several instrumented vehicles are introduced by different research groups, such as Stanford's Junior , which employs various sensors with different modalities for perceiving external and internal variables. Boss won the DARPA Urban Challenge with an abundance of sensors. RobotCar is a cheaper research platform aimed at data collection. In addition, different levels of driving automation have been introduced by the industry; Tesla's Autopilot and Google's self-driving car are some examples. However, finding the coordinate of the object in the frame has become challenging due to various factors such as variations in viewpoints, poses, scales, lighting, occlusions, etc. After the development of the deep neural network, computer vision gained even more strides. With the advances in sensing and computational technologies in the field of computer vision, the performance of traditional manual feature-based object detection algorithms has been compared to that of the deep learning-based object detection algorithms because of continuous growth in large volumes of data and fast development of hardware, particularly Multicore Processors and Graphical Processing Units (GPUs). Furthermore, the deep learning-based algorithms exceed the traditional algorithms in terms of detection speed and accuracy. The deep learning methods have gained much attention due to the promising results it has achieved in multiple fields, such as image classification , segmentation, and moving object detection and tracking. Object counting, overtaking detection, object classification, lane change detection, emergency vehicle detection, traffic control, traffic sign, light identification, license plate recognition, and many other applications of deep learning-based detection can be found in every Intelligent Transportation System (ITS) field. Object detection has been a hot topic for many researchers over the past decade. In the following literature: deep learning-based object detection, on-road vehicle detection, object detection, and safe navigation by Markov random field (MRF) in which authors aim to analyze the deep learning-based algorithms without considering recent improvements in the deep learning field.

II. LITERATURE REVIEW

SN	Title	Author	Dataset	Methodology	Findings
1	3D Object Detection with SLS-Fusion Network in Foggy Weather Conditions (2021)	Nguyen Anh Minh Mai, Pierre Duthon, Louahdi Khoudour, Alain Crouzil, Sergio A. Velastin	KITTI, Multi Fog KITTI datasets	SLS Fusion approach is used.	After the improvement, performance is increased by 42.67%.
2	Road Object Detection: A Comparative Study of Deep Learning-Based Algorithms(2021)	Malik Haris, Adam Glowacz	KITTI dataset	Algorithms used are Region-based Fully Convolutional Network (R-FCN), Mask Region-based Convolutional Neural Networks (Mask R-CNN), Single Shot Multibox Detector (SSD), RetinaNet, and You Only Look Once v4 (YOLO-v4).	This model has much lower localization errors than traditional YOLO, which results in an accuracy improvement about 11.9% MAP on KITTI dataset.
3	Machine and Deep Learning Techniques for Daytime Fog Detection in Real Time with In-Vehicle Vision Systems Using the SHRP 2 Naturalistic Driving Study Data (2021)	Md Nasim Khan, Mohamed M. Ahmed	Video data from the SHRP 2 Naturalistic Driving Study (NDS)	Support vector machine (SVM) and K-nearest neighbor (K-NN) algorithms, convolutional neural network (CNN)	Accuracy is 92% and 91% for SVM and K-NN classifiers, CNN produced much greater accuracy of 99%.
4	Adaptive Model for Object Detection in Noisy and Fast-Varying Environment (2021)	Dung Nghi Truong Cong, Louahdi Khoudour, Catherine Achard	Camera data from moving train framework of BOSS European.	Modeling each pixel intensity with a single Gaussian distribution. Mixture of Gaussians and probability density function estimated by kernel function	Achieved processing the higher-level vision tasks and algorithm for foreground object extraction in complex scenes with non-stationary background

III. DATASET DESCRIPTION

DAWN dataset: Detection in Adverse Weather Nature

To the best of our knowledge, few datasets address the problem of adverse weather conditions by certain types of synthetic weather in images plus a few real-world images. For instance, Sakaridis et al. proposed two datasets; synthetic foggy cityscapes and foggy driving datasets to investigate vehicle detection and defogging algorithms in traffic environments with 8 classes. Li et al. introduced a benchmark to evaluate deraining algorithms in the traffic scene consisting of rain in driving and surveillance datasets. This dataset consists of synthetic and real-rainy environment of 2,495 and 2,048 images, respectively. But we are only using the fog images as we are focusing on the objection in foggy whether dataset. So a few sample images of the foggy dataset is shown below:

Fig. 1: Sample images of the fog images of the DAWN dataset

There is a need for a dataset of real-world images addressing the shortcomings of the aforementioned datasets considering imaging in bad weather conditions. Currently, it is uncertain how deep learning algorithms would carry out on the wild through the influence of cross-generalization for adverse weather conditions. In addition, how the progress of these algorithms is standardized and applied safely in the ITS's applications. To this end, we introduce a novel dataset of real-world images collected under various adverse weather conditions, which we called "DAWN: Detection in Adverse Weather Nature". It is designed to support the research in ITS's applications for safety opportunities. The unique characteristics of DAWN dataset give researchers a chance to examine aspects of vehicles detection that have not been examined before in the literature as well as issues that are of key importance for autonomous vehicles technology and ITS safety applications.

The goal of DAWN dataset is to investigate the performance of vehicle detection and classification methods on a wide range of natural images for traffic scenes in the cross-generalization of adverse weather conditions, which are divided into four categories according to the weather (i.e., fog, snow, rain and sand). DAWN dataset contains significant variation in terms of vehicle category, size, orientation, pose, illumination, position and occlusion. Moreover, this dataset exhibits a systematic bias for traffic scenes during nasty winter weather, heavy snow hits, sleet rain, hazardous weather, sand and dust storms.

To ensure an accurate evaluation, the traffic scenes are comprehensive with normally moving and congested traffic, combined motorway, highway, urban roads and intersections which built up of several countries to cover the weather change of the different regions in the universe. Annotations of the vehicles are consistent, accurate and exhaustive for vehicles' classes (e.g., car, bus, truck, motorcycle and bicycle) with the presence of the human as cyclist and pedestrian. Examples of annotations in DAWN dataset are illustrated in Figure. 2.

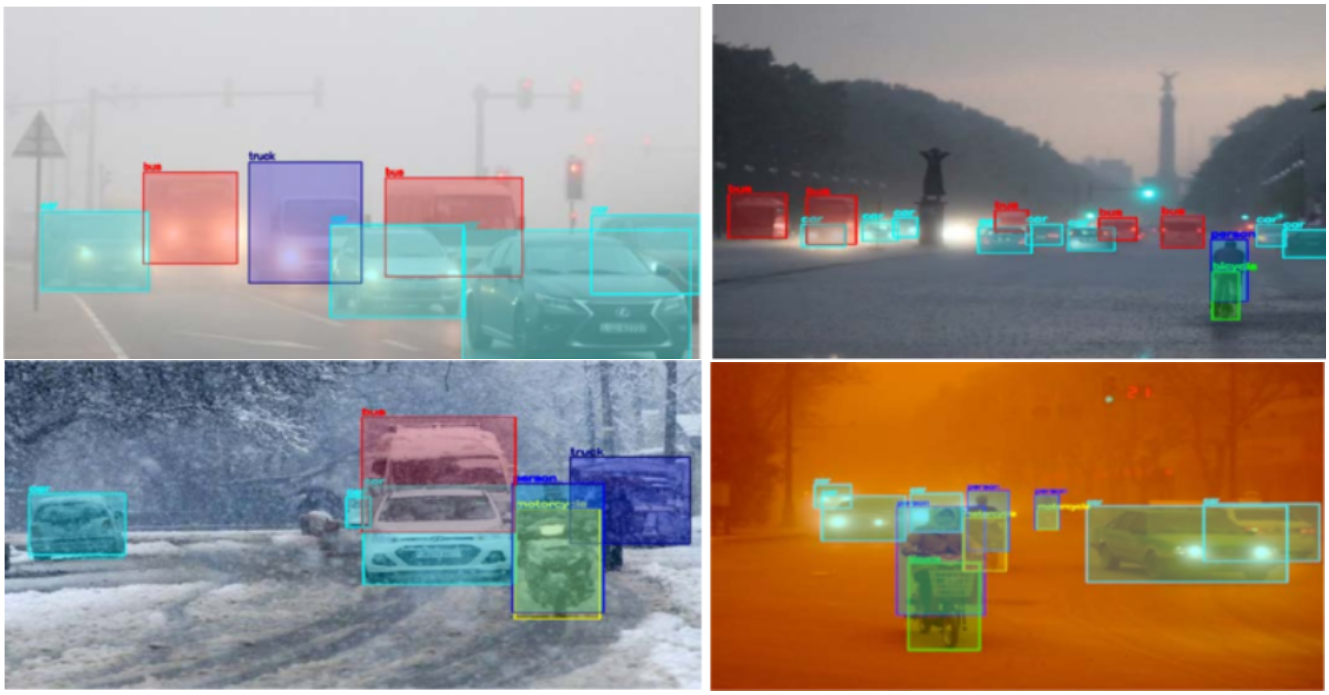


Fig. 2. Samples images of annotations in DAWN dataset

IV. METHODOLOGY

We used two state of art deep learning models like Region based Convolution Neural Network(RCNN) and You Only Look Once(Yolo-V5).

Data Augmentation:

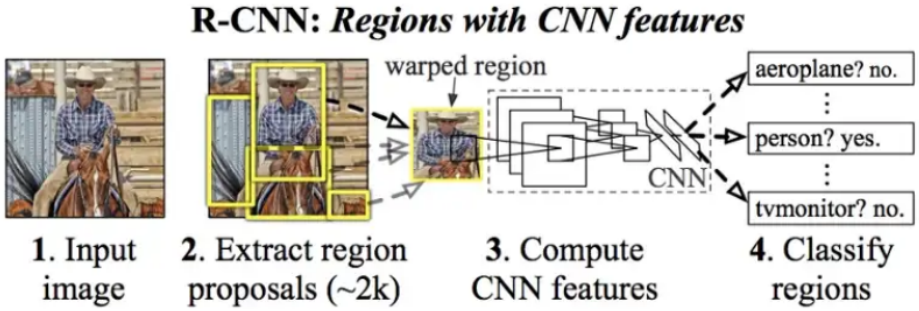
- The first preprocessing we need to do is down sampling the size of all images to 256x256.
- This is done to ensure model can be trained for all instances present in an image
- Since the annotated bounding box coordinates will also change on resizing, the coordinates should be scaled up or down using a scaling factor.
- This is a two-tuple vector denoting the ratio between the original size of the image to the rescaled image.
- The next pre-processing is doing horizontal flipping of the images.
- To adjust for the bounding boxes in such cases, the present value should be subtracted from the actual dimension of the images.

Region Based Convolution Neural Network:

Object detection consists of two separate tasks that are classification and localization. R-CNN stands for Region-based Convolutional Neural Network. The key concept behind the R-CNN series is region proposals. Region proposals are used to localize objects within an image.

Working of RCNN :

RCNN: Regions with feature



RCNN: Working Details. Source https://arxiv.org/pdf/1311.2524.pdf.

As can be seen in the image above before passing an image through a network, we need to extract region proposals or regions of interest using an algorithm such as selective search. Then, we need to resize (wrap) all the extracted crops and pass them through a network.

Finally, a network assigns a category from C + 1, including the 'background' label, categories for a given crop. Additionally, it predicts delta Xs and Ys to shape a given crop.

Extract region proposals:

Selective Search is a region proposal algorithm used for object localization that groups regions together based on their pixel intensities. So, it groups pixels based on the hierarchical grouping of similar pixels.

Positive vs. negative examples:

After we extract our region proposal, we also have to label them for training. Therefore, we label all the proposals having IOU of at least 0.5 with any of the ground-truth bounding boxes with their corresponding classes. However, all other region proposals that have an IOU of less than 0.3 are labelled as background. Thus, the rest of them are simply ignored.

Bounding-box regression:

$$t_x = (G_x - P_x)/P_w$$
$$t_y = (G_y - P_y)/P_h$$
$$t_w = \log(G_w/P_w)$$
$$t_h = \log(G_h/P_h)$$

The image above shows deltas that are to be predicted by CNN. So, x, y are centre coordinates. whereas w, h are width and height respectively. Finally, G and P stand for ground-truth bounding box and region proposal respectively. It is important to note that the bounding box loss is only calculated for positive samples.

Loss:

The total loss is calculated as the sum of classification and regression losses. However, there is a coefficient lambda. Note that the regression loss is ignored for negative examples.

Architecture:

Typically, we pass the resized crops through VGG 16 or ResNet 50 in order to get features. They are subsequently passed through fully connected layers that output predictions. There are several drawbacks to why it is not used anymore. The largest disadvantage is the selective search algorithm used for proposal extraction. Considering that the algorithm is executed on cpu, the inference time becomes slow. Additionally, all the proposals have to be resized and passed through the network, which also adds an overhead. Therefore, I am going to write about other algorithms that were introduced to overcome these problems.

Selective Search :

Selective search is a greedy algorithm that combines smaller segmented regions to generate region proposal. This algorithm takes an image as input and output generate region proposals on it. This algorithm has the advantage over random proposal generation is that it limits the number of proposals to approximately 2000 and these region proposals have a high recall.

Why Selective Search Fails here?

- Selective search technique in RCNN generates regions, based on an algorithm called hierarchical clustering: based on colours and gradients present in an image.
- However, in case of a foggy dataset, identifying them is a tough task and the algorithm does not perform very accurately.
- Thus, for finding the accurate regions of interest, a method called instance selective search that draws anchor boxes around an instance is used in our work.

Challenges of R-CNN:

- Selective Search algorithm is very rigid and there is no learning happens in that. This sometimes leads to bad region proposals generation for object detection.
- Since there are approximately many proposals. It takes a lot of time to train the network. Also, we need to train multiple steps separately (CNN architecture, SVM model, bounding box regressor). So, this makes it very slow to implement.
- R-CNN cannot be used in real time because it takes approximately 50 sec to test an image with bounding box regressor.
- Since we need to save feature maps of all the region proposals. It also increases the amount of disk memory required during training.

As RCNN shows a poor performance we are aiming to improve our performance using another deep learning method which is Yolo V5.

You Only Look Once(Yolo V5):

The YOLO family of object detection models grows ever stronger with the introduction of YOLOv5. YOLO an acronym for 'You only look once', is an object detection algorithm that divides images into a grid system. Each cell in the grid is responsible for detecting objects within itself. YOLO is one of the most famous object detection algorithms due to its speed and accuracy.

The History of YOLO:
YOLOv5
Shortly after the release of YOLOv4 Glenn Jocher introduced YOLOv5 using the Pytorch framework.
YOLOv4
With the original authors work on YOLO coming to a standstill, YOLOv4 was released by Alexey Bochoknovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao.
YOLOv3
YOLOv3 improved on the YOLOv2 paper and both Joseph Redmon and Ali Farhadi, the original authors, contributed.
YOLOv2
YOLOv2 was a joint endevor by Joseph Redmon the original author of YOLO and Ali Farhadi.
YOLOv1
YOLOv1 was released as a research paper by Joseph Redmon.

Architecture:

There are three reasons why we choose Yolov5 as our first learner. Firstly, Yolov5 incorporated cross stage partial net-work (CSPNet) into Darknet, creating CSPDarknet as its backbone. CSPNet solves the problems of repeated gradient information in large-scale backbones, and integrates the gradient changes into the feature map, thereby decreasing the parameters and FLOPS (floating-point operations per second) of model, which not only ensures the inference speed and accuracy, but also reduces the model size. In forest fire detection task, detection speed and accuracy is imperative, and compact model size also determines its inference efficiency on resource-poor edge devices. Secondly, the Yolov5 applied path aggregation network (PANet) as its neck to boost information flow. PANet adopts a new feature pyramid network (FPN) structure with enhanced bottom-up path, which improves the propagation of low-level features. At the same time, adaptive feature pooling, which links feature grid and all feature levels, is used to make useful information in each feature level propagate directly to following subnetwork. PANet improves the utilization of accurate localization signals in lower layers, which can obviously enhance the location accuracy of the object. Thirdly, the head of Yolov5, namely the Yolo layer, generates 3 different sizes (18×18, 36×36, 72×72) of feature maps to achieve multi-scale prediction, enabling the model to handle small, medium, and big objects. A forest fire usually develops from small-scale fire (ground fire) to medium-scale fire (trunk fire), then to big-scale fire (canopy fire). Multi-scale detection ensures that the model can follow size changes in the process of fire evolution. The network architecture of Yolo V5 is shown in the below figure.3.

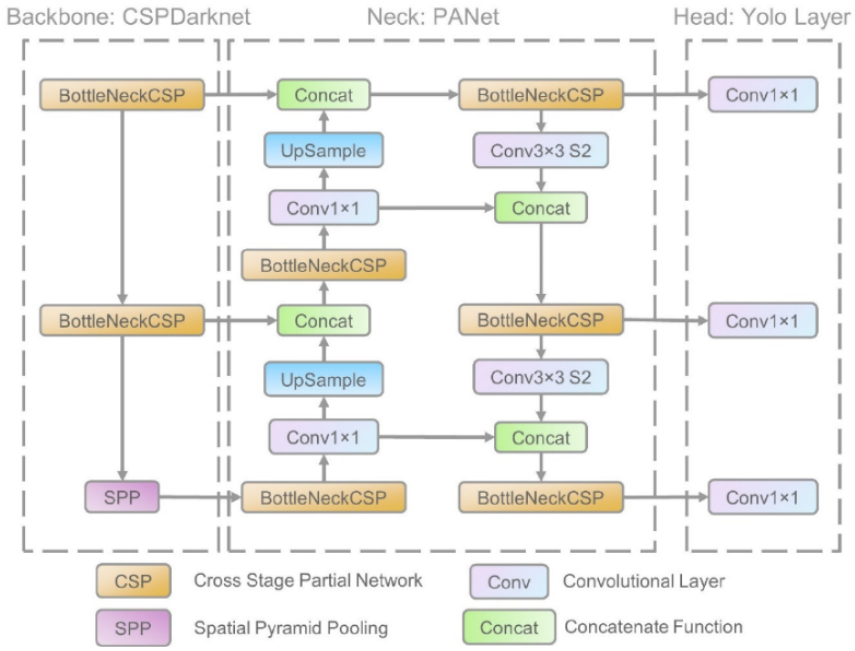


Fig. 3. Shows the network architecture of Yolo V5

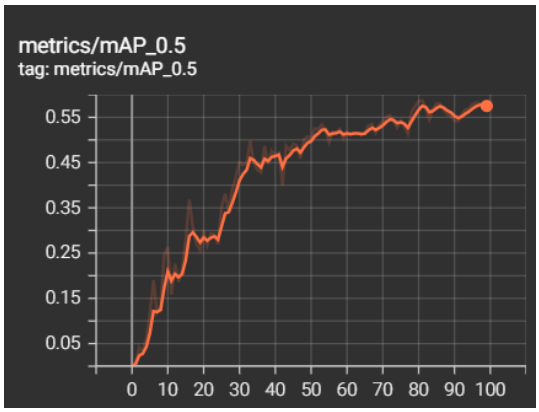
V. RESULTS

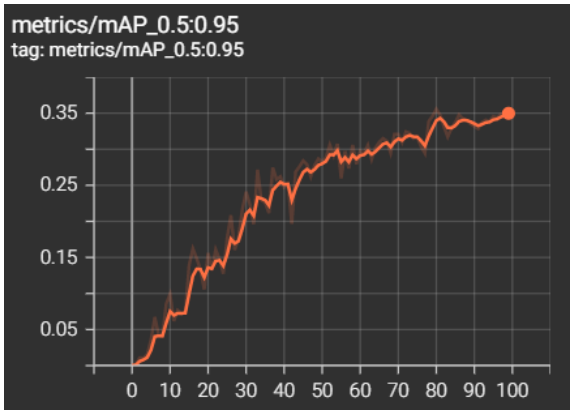
i. RCNN

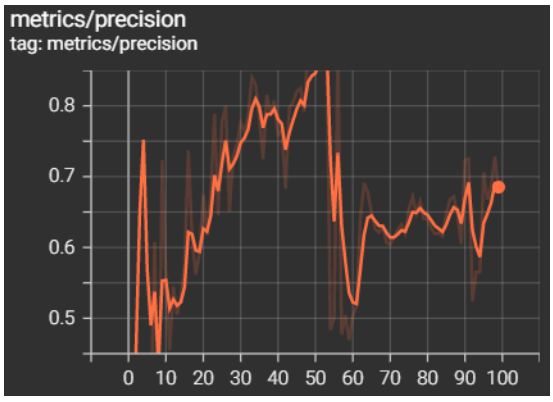
All Classes : Train mAP: 0.231
Test mAP: 0.207

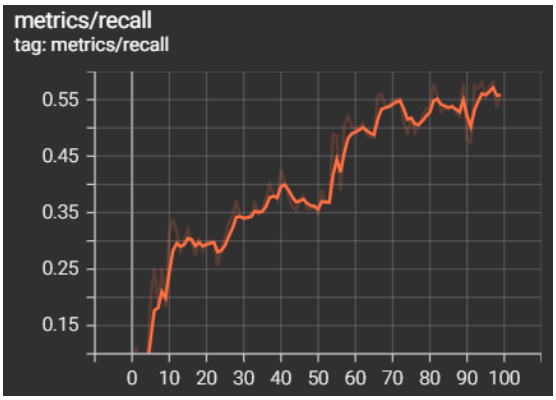
Binary Class(Car+BG): Train mAP: 0.279
Test mAP: 0.216

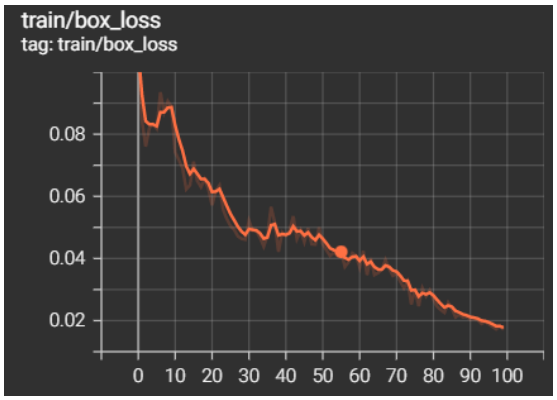
ii. YOLO v5

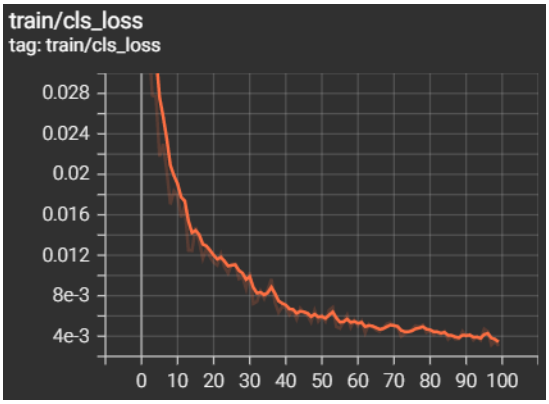


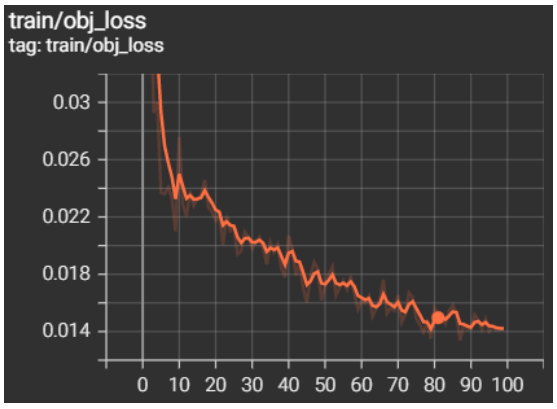


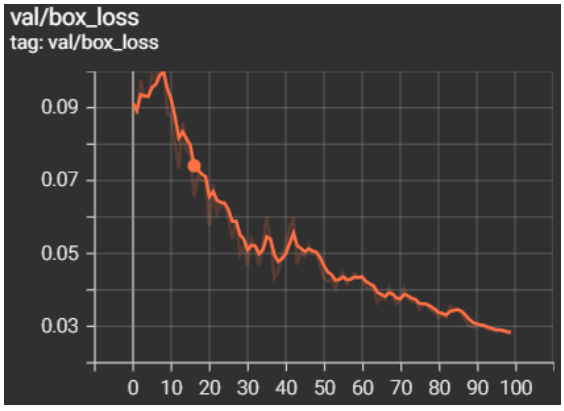


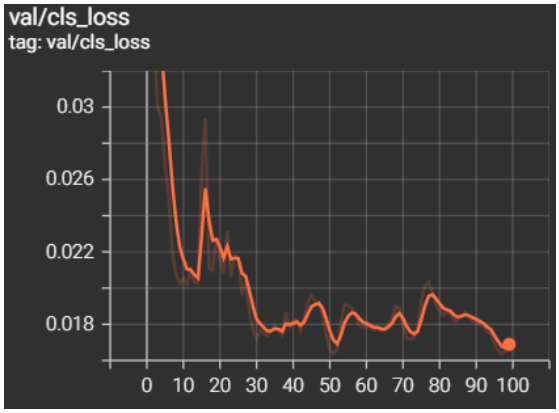


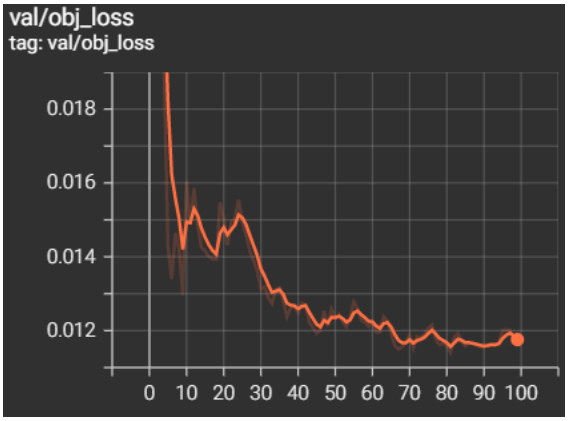












VI. CONCLUSION

There is a severe class imbalance in the dataset. Due to this, the metrics for RCNN such as classification accuracy is very less for multiclass problem. However, such a problem is optimized by YOLO v5's performance. With the current dataset without any preprocessing, YOLO v5 can give better results comparatively. So as a part of our future work, we would preprocess this dataset with the more optimized model for even more better results.

ACKNOWLEDGMENT

We would like to acknowledge our guide Mr. Sajith Variyar for proper guidance and instructions for completing this project.

REFERENCES

1. Mai NAM, Duthon P, Khoudour L, Crouzil A, Velastin SA. 3D Object Detection with SLS-Fusion Network in Foggy Weather Conditions. Sensors. 2021; 21(20):6711. https://doi.org/10.3390/s21206711

2. Truong Cong, D.N., Khoudour, L., Achard, C., Flanquart, A. (2011). Adaptive Model for Object Detection in Noisy and Fast-Varying Environment. In: Maino, G., Foresti, G.L. (eds) Image Analysis and Processing – ICIAP 2011. ICIAP 2011. Lecture Notes in Computer Science, vol 6978. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-24085-0_8

3. F. Esen, A. Degirmenci and O. Karal, "Implementation of the Object Detection Algorithm (YOLOV3) on FPGA," 2021 Innovations in Intelligent Systems and Applications Conference (ASYU), 2021, pp. 1-6, doi: 10.1109/ASYU52992.2021.9599073.

4. Khan MN, Ahmed MM. Machine and Deep Learning Techniques for Daytime Fog Detection in Real Time with In-Vehicle Vision Systems Using the SHRP 2 Naturalistic Driving Study Data. Transportation Research Record. June 2022. doi:10.1177/03611981221103236

5. K. Qian, S. Zhu, X. Zhang and L. E. Li, "Robust Multimodal Vehicle Detection in Foggy Weather Using Complementary Lidar and Radar Signals," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 444-453, doi: 10.1109/CVPR46437.2021.00051.

6. Mahaur, B., Singh, N. & Mishra, K.K. Road object detection: a comparative study of deep learning-based algorithms. Multimed Tools Appl 81, 14247–14282 (2022). https://doi.org/10.1007/s11042-022-12447-5

7. Renjie Xu, Haifeng Lin & Kangjie Lu. A Forest Fire Detection System Based on Ensemble Learning. (2021). https://www.researchgate.net/publication/349299852_A_Forest_Fire_Detection_System_Based_on_Ensemble_Learning

8. Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik. UC Berkeley Rich feature hierarchies for accurate object detection and semantic segmentation Tech report (v5). 2014 https://arxiv.org/pdf/1311.2524.pdf

AUTHORS

First Author – Rohith G (CB.EN.U4AIE19026), B. Tech – CSE-AI.
Second Author – Vasudevan KM, (CB.EN.U4AIE19067), B. Tech – CSE-AI.