# Transfer Learning Based Neural Network for Object Detection

**[1] Dr.N.R.Gayathri, [2] Mr.T.B.Dharmaraj, [3]Mrs.K.Sona, [4] Mrs.S.Vasumathikannaki,**
**[5]Dr. A.Kodieswari**

[1]AssociateProfessor, Department of Artificial Intelligence and data science, Christ The King Engineering College, Karamadai, Coimbatore.

[2]AssociateProfessor, Department of Computer Science and Engineering, Christ The King Engineering College, Karamadai, Coimbatore.

[3]AssistantProfessor, Department of Computer Science and Engineering, Christ The King Engineering College, Karamadai, Coimbatore.

[4]AssistantProfessor, Department of Computer Science and Engineering, Christ The King Engineering College, Karamadai, Coimbatore.

[5]Associate Professor, Department of Artificial Intelligence and Machine Learning, Bannari Amman Institute of Technology, Sathyamangalam.

## Abstract

Computer vision is a field of study that focuses on how computers can interpret and analyze visual information from the world around us. This includes tasks such as image and video recognition, object detection, facialrecognition and scene reconstruction.Therefore, video analysis and understandingthe images has become necessary and challenging issue. This study aimed at presenting the model based on benchmark MSCOCO datasetof image data. Highly accurate object detection-algorithms and architectures such asFaster R-CNN, Mask R-CNN with backbone architectures Resnet, Inception and ResNeXt are fast yet highly accurate ones like YOLOv3 and YOLOv4 captures both low-level and high-level features. These models exhibit different behaviors in terms of network architecture, training methods, and optimization techniques, etc. Each and every object in an image is identified by the area object in a highlighted rectangular boxes and tag is assigned to each and every object.The accuracy in detecting theobjects is checked by different parameters such as accuracy, frames per second(FPS) and mean average precision (mAP). Also, the performance of the presented model with YOLOv4 andDarknet-53as a backbone architecture for transfer learningor fine-tuned for specific computer vision taskachieves 49.2% mAP, which outperforms the baseline by 3.8%.

**Keywords: R-CNN, YOLOv3, YOLOv4, ResNeXt, CNN, SVM, HMMs, DarkNet, EfficientNet.**

## Introduction

In computer vision, pattern recognition techniques are often used to analyze visual data and extract useful features that can be used for tasks such as object detection and image classification. These techniques include machine learning algorithms such as neural networks, decision trees, and support vectormachines [1].While vision focuses specifically on visual data, pattern recognition techniques can be applied to a wide range of data types.

The goal of computer vision is to enable computers to understand and interpret the visual world in the same way that humans do [2]. This involves developing algorithms and techniques that can identify and classify objects in images and videos, recognize patterns, and perform other visual tasks.It is used in a wide range of applications, including medical imaging, robotics, autonomous vehicles [3], surveillance and entertainment. For example, computer vision is used in self-driving cars to detect and identify objectssuch as pedestrians, traffic signsand other vehicles.

The development of computer vision has been made possible by advancements in artificial intelligence and machine learning which enable computers to learn from large amounts of data and improve their performance over time. Computer vision also draws on techniques from other fields, such as mathematics, statistics and computer science to develop algorithms and models that can analyze and interpret visual data.Computer vision works by using algorithms and mathematical models to analyze and interpret visual data such as images or videos. The vision techniques can be broadly categorized into two types: traditional computer vision and deep learning-based computer vision. Traditional vision techniques use hand-crafted features and mathematical models to analyze visual data while deep learning-based computer vision techniques use neural networks to automatically learn features and patterns from data.

However, computer vision is a rapidly evolving field with many exciting applicationsand it continues to advance through research and development in machine learning, computer scienceand other related fields.There are many algorithms and mathematical models used in computer vision to perform various tasks. Here are a few examples:

Convolutional Neural Networks (CNNs): CNNs are a type of deep learning algorithm used for image and video recognition. They are based on a layered architecture that can automatically learn features from data.[10]

Support Vector Machines (SVMs): SVMs [10] are a type of machine learning algorithm used for image classification and object detection. They work by separating data into classes using a hyperplane in a high-dimensional space.

Feature-based methods: These algorithms extract features from images or videos such as edges, cornersor shapesand use them to perform tasks like object recognition or tracking.

Optical Flow: Optical flow is a technique that estimates the motion of objects in an image or video. It works by analyzing the patterns of pixel intensity changes over time.

Hidden Markov Models (HMMs): HMMs are a type of probabilistic model used for tasks such as speech recognition and gesture recognition. They model the probability distribution of sequences of observations and use this to infer the underlying hidden state.

Stereo Vision: Stereo vision is a technique used to reconstruct a 3D scene from two or more 2D images. It works by analyzing the disparities between corresponding points in the images which can be used to calculate the depth of objects in the scene.

These are just a few examples of the many algorithms and models used in computer vision. The choice of algorithm or model depends on the specific task,the type of dataand other factors such as computational resources and accuracy requirements.

The following are a summary of the contributions made in this paper:

1. Load the pre-trained model.Modify the last layer of the model for the new task.Build a new model with the modified last layer,Freeze the weights of the pre-trained layers,Compile the new model with an appropriate loss function and optimizer and Train the new model on the new task data.
2. To understand the subtle differences between the various types of images we may utilise the most current CNN models (ResNet and DarkNet)[14].

3.To detect the images we have used parameter tuning and transfer learning. The results of these experiments show that this technique is more accurate.

A CNN [13] design for detecting object in vehicle images is suggested in this paper. The rest of the manuscript is organised as follows. Section II details related works. Section III details the proposed model. The proposed architecture's experimental findings and analysis summarized in Section IV. Section V deals with the conclusion.

## Section II
## Literature Review

JINYOUNG PARK et al. (2022) proposed a model based on Mask RCNN [8], the Image Segmentation model, but was designed in a lightweight version, so that it could be used on a platform of the vessel where the use of high performing computers is limited. The lightweight Mask RCNN model showed 64% lower number of parameters compared to the base model. However, further studies are needed as the model needs improvement of recognition in low-resolution images.[9]

Long Qin et al. (2022) proposed a real-time salient object detection network named increase-decrease YOLO (ID-YOLO) to discriminate the critical objects within the drivers' fixation region.

educe the interference of irrelevant scene information, showing potential practical applications in intelligent or assisted driving systems. This work only predicts the object located in the driver-gazed location, and it cannot identify the traffic events. The samples includes video under more complicated driving environments, and more object categories (e.g., driving-related pedestrians, traffic lights, riders,buses, trucks and bicycles) [7]

QUNYING HE et al. (2022) presents a Weakly Supervised Faster RCNN (WSRC) that aims to train accurate traffic object detectors without using any box annotations, addressing the phenomenon that the traffic object detector trained by traditional method cannot be applied in the real traffic scene due to the lack of data annotation.[4]

Zhong-Qiu Zhao etal(2018) presented typical generic object detection architectures with a few modifications and helpful tips to further enhance detection performance. We also briefly examine a number of specific tasks, such as salient object detection, face detection, and pedestrian detection, as different specific detection tasks exhibit different characteristics.However for localization accuracy on small objects under partial occlusions, it is necessary to modify network architectures.

Zhonghong Ou et al(2023) proposed an AD-RCNN, an adaptive dynamic neural network, which incorporates three key advancements. A dynamic network for regional proposals to raise the calibre of regional proposals, A visual attention approach is then introduced to create regional characteristicsand Lastly, in order to improve the final detection results is proposed. Experimental results show that, in terms of mAP and frames per second, AD-RCNN can improve the accuracy of small object detection.[6]

## Section III

There is no single "best" CNN (Convolutional Neural Network) [13] for computer vision, as the choice of CNN architecture depends on the specific task and dataset being used. However, here are a few popular CNN architectures that have achieved state-of-the-art performance on various computer vision tasks:

ResNet (Residual Network): ResNet[14] is a deep CNN architecture that uses residual connections to help alleviate the vanishing gradient problem. It has achieved state-of-the-art performance on tasks such as image classification, object detection and semantic segmentation.

VGG (Visual Geometry Group): VGG is a CNN architecture with a simple, uniform structure that uses 3x3 convolutional filters. It has achieved strong performance on tasks such as image classification and object detection.

Inception: Inception is a family of CNN architectures that use multiple parallel convolutional layers of different filter sizes. It has achieved state-of-the-art performance on tasks such as image classification and object detection.

EfficientNet: EfficientNet [12] is a family of CNN architectures that use a compound scaling method to balance model size and accuracy. It has achieved state-of-the-art performance on tasks such as image classification and object detection.

ResNeXt (Residual Next): ResNeXt is a variant of ResNet that uses grouped convolutions to reduce the number of parameters and improve efficiency. It has achieved state-of-the-art performance on tasks such as image classification and object detection.

Again the choice of CNN architecture depends on the specific task and dataset being usedand it is important to experiment with different architectures and hyperparameters to find the best model.

YOLOv4 (You Only Look Once version 4)[10] is a state-of-the-art object detection algorithm developed by the computer vision research group at the University of Washington. It is an extension of the YOLOv3 algorithm with several improvements, including a larger model size, advanced data augmentation techniques and a better training strategy.One of the key improvements in YOLOv4 is the use of a larger model architecture which includes a deeper and wider neural network with more convolutional layers than YOLOv3. This large model size allows YOLOv4 to detect objects with higher accuracy and better localization.

Another important improvement is the use of advanced data augmentation techniques during training, such as cutmix and mosaic augmentationwhich helps to increase the robustness of the model to changes in lighting, occlusionand other factors that can affect object detection performance.In addition, YOLOv4[10] uses a better training strategy that includes a combination of traditional image classification loss and object detection loss, as well as a self-adaptive training method that adjusts the learning rate based on the training progress.

Overall YOLOv4 has shown significant improvements in object detection performance compared to previous versions and other state-of-the-art object detection algorithms. It has achieved state-of-the-art performance on several benchmark datasets, including the MS COCO dataset.The backbone architecture of YOLOv4 is based on the Darknet-53 network, which is a deep convolutional neural network with 53 layers. YOLOv4 also includes some additional improvements and optimizations to the network architecture to improve performance and accuracy.

The backbone architecture of YOLOv4 is designed to capture features at multiple scales and process them efficiently using residual connections and CSP blocks. These improvements help to improve the performance and accuracy of the YOLO object detection system.While both Darknet and ResNetare capable of achieving high accuracy in computer vision tasks, there are some key differences between them. Darknet is a neural network framework that provides a wide range of tools and features for developing object detection and recognition models, while ResNet is a specific CNN architecture that is designed for image classification. Darknet also uses different techniques for processing input data, such as spatial pyramid poolingwhile ResNet focuses on improving the efficiency of convolutional operations. If you need to develop an object detection or recognition model, Darknet may be a better choice due to its flexibility and high performance.

It's important to note that achieving high accuracy on visual data not only depends on the choice of the model but also on other aspects such as the quality and quantity of the training data, the data pre-processing and augmentation techniques, the hyperparameters of the training process, and the choice of the optimization algorithm.

**Darknet-53 architecture**

The Darknet architecture is primarily used in conjunction with the YOLO (You Only Look Once)[13] object detection system, which allows for real-time detection of objects in images and video streams. The YOLO system uses a single neural network to predict the bounding boxes and class probabilities for objects in an image or video frame.

The Darknet architecture consists of 53 convolutional layers and 4 max pooling layers. The convolutional layers are grouped into blocks with each block consisting of a combination of convolutional layers, batch normalization layers, and activation layers (typically leaky ReLU). The max pooling layers are used to reduce the spatial dimensions of the feature maps.

One notable feature of the Darknet architecture is its use of "route" layers which allow information from earlier layers to be combined with information from later layers. This can help improve the accuracy of the model by allowing it to capture both low-level and high-level features.

Here's a brief overview of the Darknet backbone architecture of YOLOv4

Input Layer: The input layer takes an image as input and resizes it to a fixed size.

Convolutional Layers: The convolutional layers are the building blocks of the network and they extract features from the input image. The YOLOv4 backbone network includes several convolutional layers with different filter sizes and numbers of filters to capture features at different scales.

Residual Connections: The residual connections are also known as skip connections are added between some of the convolutional layers. These connections help to preserve information from earlier layers and can improve the flow of gradients during training.

SPP Layer: The Spatial Pyramid Pooling (SPP) layer is added to the end of the convolutional layers. The SPP layer uses max pooling operations at multiple scales to extract features from different regions of the input image at different scales.

CSP Block: The Cross-Stage Partial (CSP) block is a new building block introduced in YOLOv4. The CSP block splits the input feature maps into two branches and processes them separately before concatenating them back together. This helps to reduce the number of computations required and improves the flow of information between the layers.

YOLO Layers: The YOLO layers are the output layers of the network and they predict the bounding boxes and class probabilities for each object in the input image. YOLOv4 uses a custom loss function called the YOLOv4 loss to optimize these predictions.[11]

In Darknet, the input is passed through a series of convolutional layers each followed by a batch normalization layer and a Leaky ReLU activation layer. The feature maps are then downloadingsampleusing max pooling layers.

The "route" layers allow information from earlier layers to be combined with information from later layers. This can help the model capture both low-level and high-level features. The output of the route layers is concatenated and passed through additional convolutional layers.

Finally, the global average pooling layer is used to reduce the feature maps to a single vector which is passed through a fully connected layer with a SoftMax activation function. The output of this layer is the bounding boxes and class probabilities for the detected objects.

There have been several studies comparing the performance of Darknet with other popular backbone architectures for object detection. Here are some of the key findings:

Darknet vs ResNet: In a study published in 2017, researchers compared the performance of Darknet-19 with several ResNet architectures (ResNet-18, ResNet-34, ResNet-50, and ResNet-101) on the PASCAL VOC and MS COCO datasets. They found that Darknet-19 outperformed all ResNet architectures in terms of both accuracy and speed.

Darknet vs Inception: Another study published in 2018 compared the performance of Darknet-53 with several Inception architectures (Inception-v2, Inception-v3, and Inception-ResNet-v2) on the MS COCO dataset. They found that Darknet-53 outperformed all Inception architectures in terms of both accuracy and speed.

Darknet vs EfficientNet: In a more recent study published in 2021, researchers compared the performance of Darknet-53 with several EfficientNet architectures (EfficientNet-B0, EfficientNet-B1 and EfficientNet-B2) on the COCO dataset. Theyfound that Darknet-53 outperformed all EfficientNet architectures in terms of both accuracy and speed.

Overall these studies suggest that Darknet is a high performance backbone architecture for object detection tasks, outperforming several other popular architectures in terms of both accuracy and speed.

Transfer learning is a popular technique used in deep learning where a pre-trained model is used as a starting point to train a new model for a different task. In YOLO, transfer learning is used by starting with a pre-trained model such as one trained on the COCO dataset and fine-tuned on a new dataset for a different object detection task. Fine-tuning involves adjusting the weights of the pre-trained model to fit the new data. This approach can save significant time and computational resources compared to training a model from scratch.

Transportation engineering frequently calls for categorization tasks to evaluate the vehicle characteristics based on the gathered vehicle data. Machine learning (ML) based algorithms have been an effective tool for object detection throughout the past few decades. Typically, distinct vehicle images should initially be acquired before gradually generating the ML model using those optimization strategies. Using this procedure, the model is able to detect the vehicle when given a fresh sample of vehicles. [16-21]

## SECTION IV
## Results and Discussions

Here is a comparison of the performance of Darknet and some other popular backbone architectures for object detection:

**Table 1: Performance of Darknet with Accuracy & FPS**

| **Backbone Architecture** | **#Parameters** | **Top-1 Accuracy** | **Top-5 Accuracy** | **Speed (FPS)** |
|---|---|---|---|---|
| Darknet-19 | 19.7M | 78.4% | 94.9% | 42.8 |

| | | | | |
|---|---|---|---|---|
| Darknet-53 | 41.7M | 78.7% | 94.6% | 31.8 |
| ResNet-50 | 25.6M | 76.2% | 92.8% | 28.1 |
| ResNet-101 | 44.5M | 76.9% | 93.8% | 17.3 |
| ResNeXt-152 | 45.6M | 79.1% | 93.9% | 36 |
| Inception-v4 | 42.6M | 80.2% | 95.3% | 7.8 |

Here is a comparison of the performance of Darknet and some other popular backbone architectures for object detection in MS COCO dataset:

**Table 2: Performance of Darknet with mAP& FPS**

| **Backbone Architecture** | **Dataset** | **#Parameters** | **mAP** | **Speed (FPS)** |
|---|---|---|---|---|
| Darknet-19 | COCO | 41.7M | 38.1 | 42.8 |
| Darknet-53 | COCO | 41.7M | 45.1 | 31.8 |
| ResNet-50 | COCO | 23.5M | 37.5 | 28.1 |
| ResNet-101 | COCO | 42.6M | 39.0 | 17.3 |
| ResNeXt152 | COCO | 45.6M | 42 | 36 |
| Inception-v4 | COCO | 45.9M | 38.2 | 7.8 |

As you can see, Darknet-53 achieves competitive performance in terms of mean average precision (mAP) and speed compared to other popular backbone architectures like ResNet-50 and ResNet-101. Inception-v4 achieves similar mAP to Darknet-53 but it is slower in terms of FPS. EfficientNet-B7 achieves the highest mAP among all the architectures listed here but it is much slower in terms of FPS compared to Darknet.

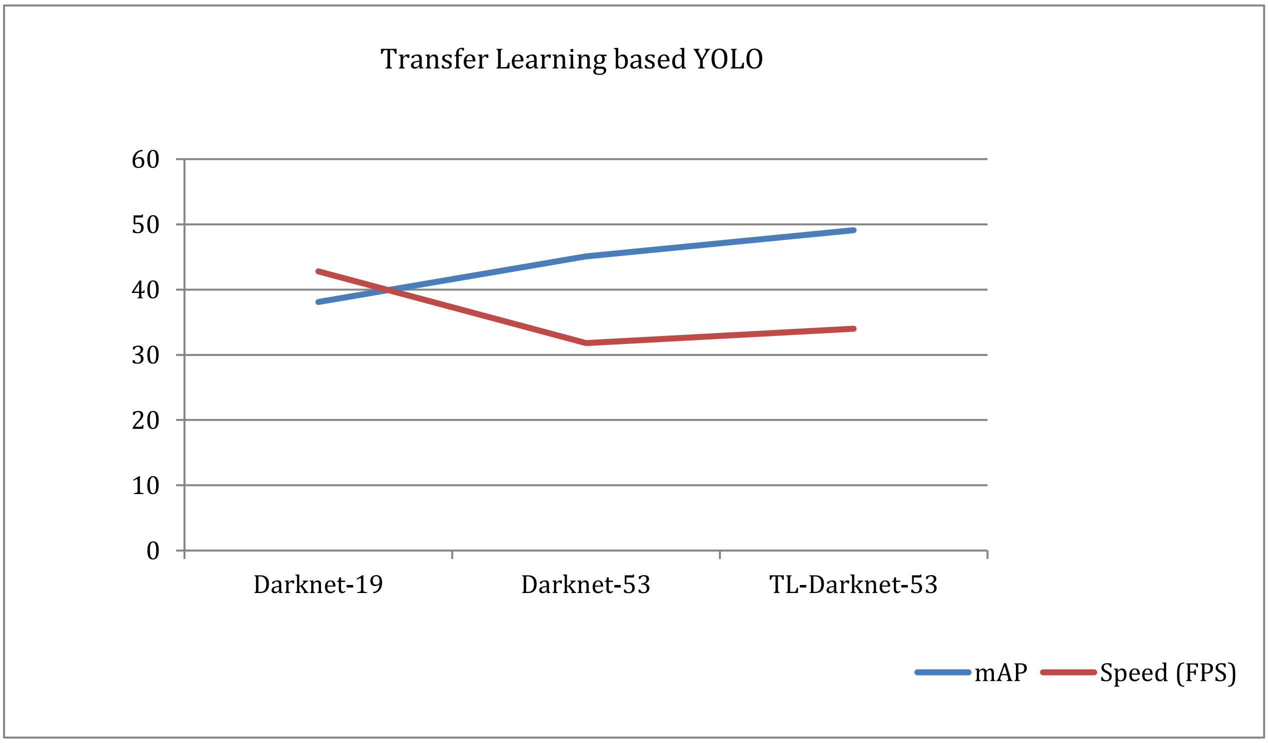

**Figure 1: Performance Comparison of Tl-Yolo5 Versus Yolo5 For Vehicle Dataset**

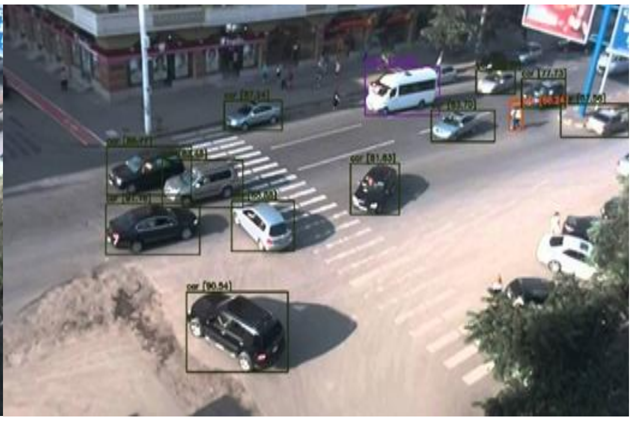

**Figure 2: Source ImageFigure 3: Object Detected Image**

## Conclusion

We developed and evaluated a novel machine learning algorithm for computer vision and image recognition tasks with a goal to achieve substantialcomputational efficiencies on computing hardware platforms. Specifically, we proposed a Transfer Learning based convolutional neural network. We analyzed and benchmarked this model against a numberof alternative models using the standard MS COCO dataset and then the new model trained is analyzed on the vehicle dataset generated.It is worth noting that Darknet-53 is a very powerful and versatile architecture and it can be combined with other techniques like data augmentation and hyper parameter optimization to achieve state-of-the-art performance on a wide range of computer vision tasks.

## References

[1] JiajiaLi , Jie Chen , Bin Sheng, Ping Li, Po Yang, David Dagan Feng and Jun Qi, “Automatic Detection and Classification System of Domestic Waste via Multimodel Cascaded Convolutional Neural Network”, 2022, VOL. 18, NO. 1

[2] .Nafiseh Zarei1, Payman Moallem, And Mohammadreza Shams2, “Fast-Yolo-Rec:Incorporating Yolo-Base Detection and Recurrent-Base Prediction Networks for FastVehicle Detection in Consecutive Images”, 2022,Digital Object Identifier,10.1109/ACCESS.2022.3221942.

[3] GuofaLi ,Zefeng Ji, Xingda Qu, Rui Zhou , and Dongpu Cao,” Cross-Domain Object Detection for Autonomous Driving: A Stepwise Domain Adaptative YOLO Approach”,2022, VOL. 7, NO.3.

[4] Qunying He 1, Jingjing Liu2, And Zhicheng Huang3, ”WSRC: Weakly Supervised Faster RCNN Toward Accurate Traffic Object Detection”,2023, Digital Object Identifier 10.1109/ACCESS.2022.3231293.

[5] . Yunyun Song ,Zhengyu Xie, Xinwei Wang, and Yingquan Zou,”MS-YOLO: Object Detection Based on YOLOv5 Optimized Fusion Millimeter-Wave Radar and Machine Vision”,2022, VOL. 22, NO. 15.

[6] Zhonghong Ou, ZhaofengnianWang ,Fenrui Xiao , Baiqiao Xiong , Hongxing Zhang, Meina Song ,Yan Zheng , and Pan Hui,”AD-RCNN: Adaptive Dynamic Neural Network for Small Object Detection”,2023, VOL. 10, NO. 5.

[7] Long Qin , Yi Shi, Yahui He, Junrui Zhang, Xianshi Zhang, Yongjie Li, Tao Deng and Hongmei Yan,”ID-YOLO: Real-Time Salient Object Detection Based on the Driver’s Fixation Region”,2022, VOL. 23, NO. 9.

[8] . Hai Wang, Zhiyu Chen, Yingfeng Cai, Long Chen ,Yicheng Li, Miguel Angel Sotelo and Zhixiong Li, ”Voxel-RCNN-Complex: An Effective 3-D Point Cloud Object Detector for Complex Traffic Conditions,”2022, VOL. 71.

[9] Jinyoung Park And Hoseok Moon,”Lightweight Mask RCNN for Warship Detection and

[10] Segmentation”,2022, Digital Object Identifier 10.1109/ACCESS.2022.3149297.

[11] . .Anima Pramanik, Sankar K. Pal, J. Maiti , and Pabitra Mitra, ”Granulated RCNN and Multi-Class Deep SORT for Multi-Object Detection and Tracking”, 2022,VOL. 6, NO.

1.

[12] Rachel Huang, Jonathan Pedoeem, Cuixian Chen, “YOLO-LITE: A Real-Time Object Detection Algorithm Optimized for Non-GPU Computers”, 2018, IEEE International Conference on Big Data (Big Data).

[13] Mingxing Tan Ruoming Pang Quoc V. Le Google Research, ”EfficientDet: Scalable and Efficient Object Detection”, 2020, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[14] . Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi, “You Only Look Once: Unified, Real-Time Object Detection”, 2016 IEEE Conference on Computer Vision and Pattern Recognition.

[15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun,”Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”, DOI10.1109/TPAMI.2016.2577031, IEEE Transactions on Pattern Analysis and Machine Intelligence.

[16] . Ibrahim H. El-Shal, Omar M. Fahmy And Mustafa A. Elattar,” License Plate Image Analysis Empowered by Generative Adversarial Neural Networks (GANs)”,2022, Digital Object Identifier 10.1109/ACCESS.2022.3157714

[17] Anwesh Kabiraj 1,2 &Debojyoti Pal 1,3 & Debayan Ganguly3 &Kingshuk Chatterjee4 & Sudipta Roy,” Number plate recognition from enhanced super-resolution using generative adversarial network”,2022, https://doi.org/10.1007/s11042-022-14018-0.

[18] Ibtissam Slimani, AbdelmoghitZaarane, Wahban Al Okaishi, Issam Atouf, Abdellatif Hamdoun, “An automated license plate detection and recognition system based on wavelet decomposition and CNN”, www.elsevier.com/journals/array/2590-0056/open-access-journal

[19] . Abdelsalam Hamdi, Yee Kit Chan ∗, Voon Chet Koo, “A New Image Enhancement and Super Resolution technique for license plate recognition”, www.cell.com/heliyon

[20] .Anmol Pattanaik1 · Rakesh Chandra Balabantaray, “Enhancement of license plate recognition performance using Xception with Mish activation function”,2022, https://doi.org/10.1007/s11042-022-13922-9.

[21] Farheen Ali, Himanshu Rathor, Wasim Akram, “License Plate Recognition System”, 2021, (ICACITE).

[22] Byung-Gil Han , Jong Taek Lee , Kil-Taek Lim and Doo-Hyun Choi, “License Plate Image Generation using Generative Adversarial Networks for End-To-End License Plate Character Recognition from a Small Set of Real Images”, 2020, applied sciences.