International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048
Volume 7, Issue 2, pp 16-23, February-2023
https://doi.org/10.47001/IRJIET/2023.702002

A Hybrid Architecture Based on Deep Learning for Object Recognition for Autonomous Driving

[1]Azhar Ali Agro Mughal, [2]Dr. Sanam Narejo, [3]Dr. Shahnawaz Talpur, [4]Ali Hasnain, [5]Muazam Ali

[1,2,3,4,5]Mehran University of Engineering & Technology, Jamshoro, Pakistan
Authors E-mail: [1]azhar697@gmail.com, [2]sanamnarejo@hotmail.com, [3]mirshan35@hotmail.com, [4]alihasnain2k19@gmail.com, [5]moazzam.ali.alvi.shah@gmail.com

Abstract **- Autonomous Vehicle (AV) is the future of Auto industry, which integrates many high-end technologies. Autonomous Vehicle (AV) will enable the road transportation fully human-independent. In the context of driverless vehicles, smart driving assistance, and cutting-edge traffic assessment, object detection plays a crucial role. Real-time accurate object identification is crucial for traffic assessment and smart driving assistance. The system's primary duty is to provide the driver or controller with a precise understanding of the road or the area around the vehicle. Though visual based autonomous vehicles have demonstrated excellent prospects, there are few problems on how to find and interpret the difficult traffic conditions of the gathered statistics. Autonomous driving has been composed of a large number of different functions individually, such as vision based object detection. Further, a vision based object detection system is divided into dynamic and stationary objects on the road. In this study, a visual based system is intended to be designed for identification of different objects. The main contributions of this research are to detect multiple objects (both dynamic and stationary) that contribute to high and low risk of collision of vehicles on roads. Experimental results showed that our both trained models achieved higher precision compared to other state-of-the-art models.**

Keywords: Deep Learning, YoloV4, Object Detection, CNN, Autonomous Vehicles.

I. INTRODUCTION

Autonomous Vehicles (AVs), also called Self-Driving cars are type of vehicles that have the ability to perceive its environment and navigate securely with little to no assistance from humans. Autonomous Vehicles integrate diverse sensing devices that perceive vehicles surroundings. The sensing devices are Radar, GPS, Odometer, Sonar and Lidar. Autonomous Vehicles controlling systems interpret the information from sensing devices to find out appropriate paths for vehicle navigation, with relevant traffic signs, while avoiding obstacles.

There has been a lot funding invested for research in both academia and industries by technology firms in recent years. A few of the technology companies where research on Autonomous Vehicles is ongoing are Tesla, General Motors, Waymo, NVIDIA, Delphi, BMW, Ford etc [1].Autonomous Driving, Driving Assistant and Traffic Analysis System [2][3] should have precise knowledge about the condition of what is ahead and near the Autonomous Vehicle [4][5].Current techniques for the visual based object detection and recognition based techniques have not evolved into their mature form because of various factors, essentially variableness in the shapes and sizes of the vehicles, jumbled environment and lightening conditions[6][7]. For Autonomous Vehicles most threatening objects are all types of vehicles and pedestrians on the road, which may cause a crash with the host vehicle [8].Deep Learning has demonstrated huge potential in recent years in adequately achieving the task of object detection and recognition [9].

Advancement has been made in the techniques of object detection and recognition for Autonomous Vehicles, there still exists risk of crash, as vehicles are surrounded by all sorts of dynamic objects (vehicles, pedestrians) and static objects (road signs, traffic lights).It is paramount to timely detect high risk (dynamic) and low risk (static) objects for real-time Autonomous Driving in different volumes of traffic and visibility conditions.

II. LITERATURE SURVEY

A) Overview of previous work

Regarding the overall issue of object detection, deep neural network-based methods have recently produced astounding results. One-step approaches (such SSD and YOLO) and two-step approaches (like Fast/Faster R-CNN, FPN, and Mask R-CNN) can be used to detect objects. Overall, one-step procedures are quicker and more effective than two-step methods in terms of performance and detection accuracy [13]. Even while methods like Faster R-CNN have good accuracy, real-time applications, use a lot of processing resources and operate slowly. [10]. One-step approaches immediately forecast bounding boxes and confidence values

© 2023-2017 IRJIET All Rights Reserved

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 2, pp 16-23, February-2023

https://doi.org/10.47001/IRJIET/2023.702002

without the need for the first step, typically referred to as Proposal [12]. Typically, the mean average precision is used to measure the detection performance.

B) Object Detection

A number of researchers in the field of autonomous driving have recently become interested in the problem of multi-object detection. To handle object detection as a regression problem, the one-stage YOLO approach was initially proposed. The state-of-the-art work YOLO can detect items quickly and reliably, but the number of objects that can be predicted is constrained by the model's spatial constraints [11]. The single shot Multi-box detection (SSD) is the name of another one-stage approach [20]. The SSD model can reach 59 FPS and 74.3% mean Average Precision (mAP) on the PASCAL VOC dataset for a 300 x 300 input size, which is significantly better than the real-time YOLO [11]. Additionally, an integrated network was used for object detection. The method's processing time is slower than the YOLO, according to experiments [11], but thanks to an improved gripping mechanism, it performed better in mAP [10]. The two-stage approaches can produce more precise detection results than the majority of one-stage methods, but the detection speed is slower.

Large activity recognition, phrase classification, text recognition, face recognition, object detection and location, picture characterisation, and other tasks can be performed with convolutional neural networks (CNN).They are composed of neurons, and each neuron has a bias and weight that may be learned. Along with an input layer, an output layer, and numerous hidden layers, it also has convolutional, pooling, fully connected (FC), and several normalising layers. To combine two pieces of information, a convolutional layer executes a convolution technique. It mimics how a single neuron responds to visual stimuli. By linking the output of a neuron cluster at one layer with a single neuron, the pooling layer reduces the dimensionality. Every neuron in one layer is linked to every other layer by the FC layer. The input images are classified into various classes using the training datasets as a basis [14].

III. METHODOLOGY

A) Dataset

MS COCO (Microsoft Common Objects in Context) collection contains 91 categories of common objects, with more than 5,000 annotated examples in 82 of them. The collection contains 328,000 images and 2,500,000 tagged instances in total. Compared to other well-known datasets, COCO has fewer categories, but each category has more occurrences. It can be advantageous to learn complex object models with precise 2D localization. The dataset has a substantially greater instance count per category than the PASCAL VOC and SUN databases do. The quantity of labeled instances per image in our collection is another important difference from past datasets, which may help with comprehension of contextual information [15]. In this paper, the dataset containing 1,500 annotated images are used in the process. The dataset is further divided into training and testing sets. The training set contains 1,000 images and testing set contains 500 images. The image resolution is 640 x 480.

B) Object Detection

The primary function of the system is object detection, which provides bounding boxes and category probabilities for each object. In this paper, dynamic objects i.e. cars, bicycle, trucks, trains, buses and persons are trained with YOLO v4 model and static objects i.e. traffic lights and stop signs are trained with optimized CNN model. This process is illustrated in Fig.1.

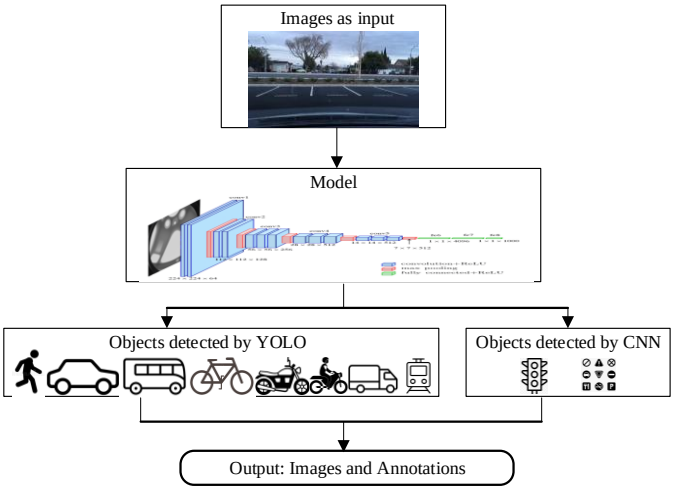

Figure 1: Block diagram of the proposed system

i) YOLO

We use YOLOv4 model architecture for object detection of dynamic objects on the road. As seen in Fig. 2, the YOLOv4 model structure is made up of three YOLO heads, CSPDarknet-53, Spatial Pyramid Pooling (SPPnet), and Path Aggregation (PAnet). CSPDarknet-53 is the core of YOLOv4, and is in charge of bringing out feature representations from the input image using 5 Resblock bodies (C1-C5). The network has 53 convolution layers, ranging in size from 1 by 1 and 3 by 3. A batch normalisation (BN) layer and a Mish activation layer connects each convolution layer. Furthermore, leaky-ReLU, which requires less processing, has been used to replace all activation functions in YOLOv4.SPPnet successfully expanded the model's receptive field using several max-pooling layers with sizes of 5, 9, and 1, while

© 2023-2017 IRJIET All Rights Reserved

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 2, pp 16-23, February-2023

https://doi.org/10.47001/IRJIET/2023.702002

PANet regularly extracted features using both top-down and bottom-up methods. To detect objects of various sizes, three YOLO heads with dimensions. As stated in (1), λ is the balance coefficient, the loss function utilised in the YOLOv4 model is composed of three parts: the object localization offset loss *Lloc*, the object confidence loss *Lconf*, and the object classification loss *Lcla*.

$$Loss = \lambda_1 L_{conf} + \lambda_2 L_{cla} + \lambda_3 L_{loc} \quad (1)$$

$$L_{conf} = -\sum (Obj_i \ln(p_i) + (1 - Obj_i)\ln(1 - p_i)) \quad (2)$$

$$L_{cla} = -\sum_{i \epsilon Box} \sum_{j \epsilon class} (O_{ij} \ln(p_{ij}) + (1 - O_{ij})\ln(1 - p_{ij})) \quad (3)$$

$$L_{loc} = 1 - IOU(A, B) + \frac{d^2_{AB}(A_{ctr}, B_{ctr})}{l2} + \alpha v \quad (4)$$

In (2), *Obji* anticipates whether an object is present in object bounding box*i*, and value of the result is either 0 or 1. The likelihood that an actual object is present in the prediction box is known as *pi*. The sigmoid function's value is calculated to get the probability value. In (3), the terms denotes respectively *Oij* and *pij* whether *j*-class object and probability in the prediction boundary box *i* exists. YOLO v4 performance is measured the Complete Intersection Over Union (CIOU) algorithm [16], equation (4) can be used to determine the object localization offset loss. Here, the aspect ratio αv and Euclidean distance of the center point (A_{ctr}, B_{ctr} for the anticipated bounding box A and the Ground Truth bounding boxes B are calculated.

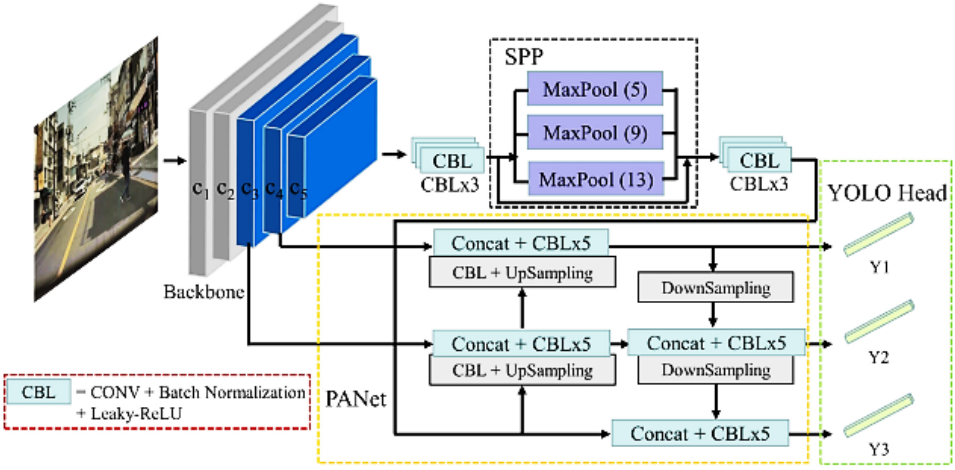

Figure 2: The model structure of YOLOv4

ii) CNN

In this study, Tensor flow's Sequential model was initiated. Then four Convolutional layers were added using a 'for' loop, with the number of output filters taking on increasing values, and the kernel of height and width three for the convolutional window, a default stride (1,1), and a ridge regression penalization. Along with each of the layers, a batch normalization step was added which further increased the accuracy. The transformation used in this phase keeps the output standard deviation and mean both near to 1.

ReLu activation function was chosen, as was the average pooling layer with a pool size of (2,2), halving the input in both spatial dimensions. This is where the loop ends, and the model proceeds with a Dropout of 0.2 which drops 20% of the input units at each update, which helps prevent overfitting. After that, the input was flattened, Activation function ReLu was chosen for the next layer and the dropout was again set to 20%. Another Dense Layer was added with as many nodes as there are classes and a ridge regression bias regularizer equal to 0.01. Finally, the softmax activation function was chosen and as an Optimizer Adam was selected. This model is run for 50 epochs.

C) Experimental Results

In this section, results of the experiments are presented. We conducted our experiments in two steps. Firstly, we trained YOLOv4 for detecting high-risk dynamic objects on the roads, and secondly CNN was used to detect low-risk static objects.

YOLOv4 has performed really well in detecting dynamic objects on the road. Our suggested system can analyse rather effectively in a variety of conditions, such as at night, in the day, or with high, medium, or low traffic. While reviewing our findings, we found some encouraging trends, such as the nearly identical effectiveness with which we could detect objects in low light or throughout the day (Fig. 3). In all of the aforementioned situations, our network successfully detects all forms of small-largeitems, even when there is moderate to low traffic on the road (Fig. 4). Even when things are far away or there is heavy traffic on the road, our suggested neural network can recognise them well (Fig. 5).

© 2023-2017 IRJIET All Rights Reserved

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 2, pp 16-23, February-2023

https://doi.org/10.47001/IRJIET/2023.702002

Figure 3: Objects detection in day and night (low) light

Figure 4: Object detection in moderate and low traffic

Figure 5: Object detection in voluminous traffic

Our optimized CNN model has also performed really well in detecting the static objects (Fig. 6).

© 2023-2017 IRJIET All Rights Reserved

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 2, pp 16-23, February-2023

https://doi.org/10.47001/IRJIET/2023.702002

IV. CONCLUSION

In this study, our suggested approach can analyse in a variety of conditions, such as night, day, heavy, medium, or low traffic, fairly effectively. Table I lists the mAPs based on IoU for the trained YOLO model. Table II lists the mAPs based on IoU for the proposed CNN model. The learning, testing accuracy and learning and testing loss of our proposed CNN model is shown in Fig. 7. Our trained and YOLO and CNN model achieved a mean Average Precision (mAP) of 55.5 and 58.57 respectively.

Figure 6: Low risk stationary objects detection

Table I: Performance and Accuracy time of YOLO model

Task	Performance		FPS	Input size
Object Detection	APs	21.53	63.5	640 x 480
	AP*m*	43.54		
	AP*l*	59.67		
	AP.*50*	68.31		
	AP.*75*	49.27		
	AP.*95*	37.34		
	mAP	55.5		

Table II: Performance and Accuracy time of YOLO model

Task	Performance		FPS	Input size
Object Detection	APs	24.5	53.2	640 x 480
	AP*m*	47.11		
	AP*l*	63.92		
	AP.*50*	72.31		
	AP.*75*	54.63		
	AP.*95*	42.38		
	mAP	58.57		

© 2023-2017 IRJIET All Rights Reserved

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 7, Issue 2, pp 16-23, February-2023
https://doi.org/10.47001/IRJIET/2023.702002

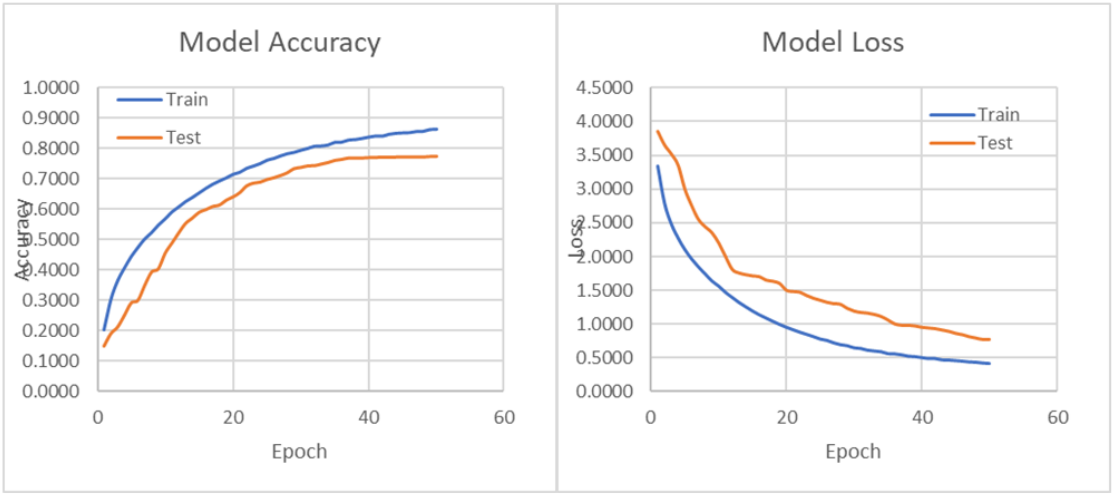

Figure 7: The learning, testing accuracy and learning and testing loss of our proposed CNN model

We evaluated certain state-of-the-art techniques and their accuracy with our suggested models in Table III.

Table III: The comparison of speed and accuracy YOLO and CNN model [20, 21]

S. No.	Method	Backbone	APs	APm	APl	FPS
1	SSD	VGG-16	6.14	24.85	40.52	25.2
2	YOLOv3	Darknet-53	10.62	29.36	42.17	32.5
3	PP-YOLO	ResNet50	19.76	41.89	54.92	68.5
4	YOLOv3-SPP	Darknet-53	19.52	35.13	45.52	27.5
5	FE-YOLO	Darknet-53	20.83	42.69	55.87	67.4
6	Faster R-CNN	ResNet101	15.6	38.7	50.9	42.61
7	Faster R-CNN w FPN	ResNet101	18.2	39.0	48.2	41.22
8	Faster R-CNN w TDM	ResNet101	16.2	39.8	52.1	52.11
9	proposed YOLO	Darknet-53	21.53	43.54	59.67	63.5
10	proposed CNN	Sequential	24.50	47.11	63.92	53.2

V. FUTURE SCOPE

We have demonstrated how to use the model to detect, classify objects in real time. Future development should focus more on enhancing the proposed framework's overall speed. In the future study, a different pipeline capable to effectively utilize processing of single-frame and multiple-frame recognition that can be used to enhance the system's performance. Distance forecasting ought to be included in this framework because it is crucial to know how far autonomous cars are from other objects.

REFERENCES

[1] R. Kulkarni, S. Dhavalikar and S. Bangar, “Traffic Light Detection and Recognition for Self Driving Cars Using Deep Learning," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2018, pp. 1-4.

[2] Heaton, Jeff., “Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep learning.” Genetic Programming and Evolvable Machines 19, no. 1-2 (2017), 305-307.

[3] Corovic, Aleksa, Velibor Ilic, Sinisa Duric, Malisa Marijan, and Bogdan Pavkovic. “The Real-Time

© 2023-2017 IRJIET All Rights Reserved

ISSN (online): 2581-3048
Volume 7, Issue 2, pp 16-23, February-2023
https://doi.org/10.47001/IRJIET/2023.702002

Detection of Traffic Participants Using YOLO Algorithm." 2018 26th Telecommunications Forum (TELFOR), 2018.

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

[5] A.Datta, T. Islam Meghla, T. Khatun, M. Hasan Bhuiya, S. Rahman Shuvo and M. Mahfujur Rahman, "Road Object Detection in Bangladesh using Faster R-CNN: A Deep Learning Approach," 2020 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE), 2020.

[6] A.Mukhtar, L. Xia and T.B. Tang, "Vehicle detection techniques for collision avoidance systems: A review", IEEE Transactions of Intelligent Transportation Systems, Vol. 16, No. 5, Oct. 2015.

[7] S. Sivaraman and M.M. Trivedi, "Looking at vehicles on the road: A survey of vision-based vehicle detection, tracking, and behavior analysis", IEEE Transactions on Intelligent Transportation Systems, Vol. 14, No. 4, Dec. 2013.

[8] Y. Chen, D. Zhao, L. Lv and Q. Zhang, "Multi-task learning for dangerous object detection in autonomous driving", Inf. Sci., vol. 432, pp. 559-571, Mar. 2018.

[9] I.Arel, D. C. Rose and T. P. Karnowski, "Deep Machine Learning – A New Frontier in Artificial Intelligence Research [Research Frontier],"IEEE Computational Intelligence Magazine, Vol. 5, No. 4, pp. 13-18, Nov. 2010.

[10] S. H. Naghavi, C. Avaznia and H. Talebi, "Integrated real-time object detection for self-driving vehicles", Proc. 10th Iranian Conf. Mach. Vis. Image Process. (MVIP), pp. 154-158, Nov. 2017.

[11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, ''You only look once: Unified, real-time object detection,'' in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 779–788.

[12] J. Janai, F. Güney, A. Behl, and A Geiger, "Computer vision for autonomous vehicles: Problems, datasets and state-of-the art", arXiv preprint arXiv:1704.05519, 2017.

[13] Q. Zhao, T. Sheng, Y. Wang, F. Ni, L. Cai, "CFENet: An Accurate and Efficient Single-Shot Object Detector for Autonomous Driving", 2018. , [online] Available: https://arxiv.org/pdf/1806.09790.pdf

[14] Dhillon, A. and Verma, G.K. (2019). Convolutional neural network: a review of models, methodologies and applications to object detection. Progress in Artificial Intelligence, 9(2), pp.85–112.

[15] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C.L. (2014). Microsoft COCO: Common Objects in Context. Computer Vision – ECCV 2014, [online] pp.740–755.

[16] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, ''Distance-IoU loss: Faster and better learning for bounding box regression,'' in Proc. AAAI, 2020, pp. 12993–13000.

[17] Corovic, Aleksa, Velibor Ilic, Sinisa Duric, Malisa Marijan, and Bogdan Pavkovic. "The Real-Time Detection of Traffic Participants Using YOLO Algorithm." 2018 26th Telecommunications Forum (TELFOR), 2018.

[18] Naghavi, S. H., & Pourreza, H. (2018). Real-time object detection and classification for autonomous driving. 2018 8th International Conference on Computer and Knowledge Engineering (ICCKE).

[19] Chen, Z., Khemmar, R., Decoux, B., Atahouet, A., & Ertaud, J. (2019). Real time object detection, tracking, and distance and motion estimation based on deep learning: Application to smart mobility. 2019 Eighth International Conference on Emerging Security Technologies (EST).

[20] Xu, D.; Wu, Y. FE-YOLO: A Feature Enhancement Network for Remote Sensing Target Detection. Remote Sens. 2021, 13, 1311. https://doi.org/10.3390/rs13071311

[21] Tsung-Yi Lin, Priya Goyal, Roos Girshick, Kaiming He, Piotr Dollar, "Focal Loss for Dense Object Detection", International Conference on Computer Vision, pp.2999-3007, 2017.

[22] Ahmed, A., Hussain, G. ., & Raza, A. . (2021). Ultra-Wide Band Horseshoe Antenna for Cognitive Radio Applications. Journal of Applied Engineering & Technology (JAET), 5(1), 9-18.

[23] Radhan, A. R., Jokhio, F. A., Hussain, G., Javed, K., & Ahmed, A. (2022). Multi-Scale Pooling In Deep Neural Networks For Dense Crowd Estimation. Sukkur IBA Journal of Emerging Technologies, 5(1), 54-63.

© 2023-2017 IRJIET All Rights Reserved

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 2, pp 16-23, February-2023

https://doi.org/10.47001/IRJIET/2023.702002

Citation of this Article:

Azhar Ali Agro Mughal, Dr. Sanam Narejo, Dr. Shahnawaz Talpur, Ali Hasnain, Muazam Ali, "A Hybrid Architecture Based on Deep Learning for Object Recognition for Autonomous Driving" Published in *International Research Journal of Innovations in Engineering and Technology - IRJIET*, Volume 7, Issue 2, pp 16-23, February 2023. Article DOI https://doi.org/10.47001/IRJIET/2023.702002

****

© 2023-2017 IRJIET All Rights Reserved