2nd International Conference of Information Technology to enhance E-learning and other Application, (IT-ELA2021),
Baghdad college of economic sciences University, Baghdad, Iraq

Pedestrian and Objects Detection by Using Learning Complexity-Aware Cascades

Mohammed F. Alrifaie
Basra University College of Science and Technology
Basra, Iraq
m_alrifaie@yahoo.com

Omar Ayad Ismael
College of Islamic Sciences
University of Diyala
Diyala,Iraq
omarismail@uodiyala.edu.iq

Asaad Shakir Hameed
Department of Mathematics, General Directorate of Thi-Qar Education, Ministry of education
Thi-Qar, Iraq
asaadutem@yahoo.com

Mustafa B. Mahmood
Computer Department
College of Science
University of Baghdad,
Baghdad, Iraq
mostafa.bassem@sc.uobaghdad.edu.iq

2021 2nd Information Technology To Enhance e-learning and Other Application (IT-ELA) | 978-1-6654-0952-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/IT-ELA52201.2021.9773589

***Abstract*— Due to the considerable technological development in all joints of life, the trend has become significant towards automating various processes in daily life, such as smart cities, the Internet of things, and cloud services. One of the most crucial applications is self-driving cars, which will be a quantum leap in this field. The main problem with these vehicles will be how to provide the necessary accuracy to deal with various situations, such as sudden stops and pedestrian crossing. In this paper, we propose an effective method for automating autonomous vehicles by improving their ability to make appropriate decisions at the right time. For this, we rely on sequential training that is aware of the complexity. The system is trained and provided to the vehicles, where the presence of pedestrians is detected using machine learning algorithms, such as a deep convolutional neural network (CNN). The findings obtained in this research suggest a clear improvement in the vehicle's ability to make decisions and a great speed in responding to the event and parking the vehicle when passing by.**

Keywords*— complexity constrained learning, detector cascades, boosting, Real-time pedestrian detection

I. INTRODUCTION

Smart vehicles, or autonomous vehicles, are one of the most important industrial applications of the automation process. The history of autonomous vehicles goes back two or three decades, when thinking began about the possibility of dispensing with drivers [1]. The research began by conducting multiple experiments, which are represented by models that go according to a specific path, and then it has been developed until it has became able to make decisions by relying on its training with different algorithms [2].

The main problem lies in such vehicles with accuracy and speed in decision-making, as they affect people's lives directly. Thus, the error will lead to severe consequences. In fact, autonomous vehicles have a long history and many consecutive applications in this field [3]. We rely on cameras that capture images processed and corrected by dual classifiers to discover objects in the vicinity of the vehicle. There are alternative algorithms that are used to discover objects, such as the object suggestion mechanism [4]. These algorithms depend on providing the system with a dataset to be compared with the objects where this object is suggested depending on the dataset. These methods can be good at speeding the discovery of some objects compared to the data set, while one of the most important disadvantages of these algorithms is the challenges faced in detecting some unregistered objects in the data set. Also, some of most of these algorithms have many requirements, such as graphics processing units, which are a key factor in discovering objects and determining their type, but this may affect the speed. In practical applications, a trade-off must be made and a balance must be struck between accuracy and execution speed. Therefore, in some autonomous aircraft (known as drones), the GPU may not be relied upon, but rather the CPU is satisfied [5].

The basis for the work of autonomous vehicles is to distinguish and discover objects as a first step, as this is done through sensors and responsive cameras to make decisions in the shortest possible time. The next step is to determine the captured image type, whether it contains spots or not, as the spots in the image may represent things. If the image is devoid of any spots, the road is clear, and there are no problems, and this is rare. As all the objects on the side of the road, such as trees, poles, lighting and traffic signals. represent spots and are discovered in the image analysis stage. After detecting the spots, it is distinguished whether they are fixed or moving objects. If it is fixed, then the dimensions and distance are measured to determine the safety zone. However, if it is mobile, it is dealt with and distinguished through artificial intelligence algorithms, such as the CNN algorithm, that trains the system and analyzes the images to discover pedestrians, and based on periodic results that are issued in seconds [6]. The appropriate decision is taken, such as standing, turning, or slowing down. There are several methods commonly used for this purpose. For example, a series of non-consecutive classifications are used to generate waterfalls. These cascades can be relied upon if they depend on the complexity of equivalent features (having a similar complexity) [7]. However, these methods suffer from clear problems when analyzing chains of asymmetric complexity, as these methods face difficulties in understanding features of varying complexity on a large scale. This is a problem for applications requiring various feature sets [8].

978-1-6654-0952-0/21/$31.00 ©2021 IEEE

The structure of the subsequent sections in this paper represents the analysis of previous studies in the section on related studies, where the studies are reviewed and their weaknesses are analyzed to determine the most appropriate ways to solve these problems. After that, the most important characteristics of the CNN algorithm and the methods for applying and benefiting from them are reviewed in this paper, and then the system training method used is discussed in the COMPLEXITY-AWARE CASCADE section, where the methods and steps required for training are reviewed. Then, the Pedestrian Detection is covered and designed, then in the next section, the experiment and its results are reviewed in the Experiment section, and finally, the conclusion and further studies are discussed in the conclusion section.

II. CONTRIBUTION

This algorithm clearly explains the complexity of the variable features in cascading learning sequences. This paper provides a clear contribution to

- proposing an effective framework for pedestrian and objects detection.
- proposing a sequence-aware learning algorithm that balances accuracy and speed in decision-making.
- achieving state-of-the-art performance on CityPersons and Caltech-USA.

III. RELATED WORKS

Self-driving vehicles are one of the most important industrial applications that attract the attention of researchers and companies alike to provide the addition in this type of project. The reason is the main challenge in self-driving vehicles is to work on increasing the accuracy of decision-making because of the severe consequences [9].

Research and studies are directed towards establishing an attention network based on training the system to pay attention to people by relying on the so-called circle of attention. In this way, pedestrians can be detected using the CNN algorithm [10], whereas the mod that was proposed by Zhou and Yuan focused on detecting pedestrians in the road and its sides using the so-called Bi-box model. The system is based on dividing the body into two main parts (the head and the torso), each of which is represented by a box. Accordingly, the full body is estimated by estimating the circumference of each of the two boxes [11].

In the same way, the field of obstacle detection has been researched considerably as it is a real problem facing self-driving vehicles, given that the size of the obstacle affects the vehicle's path. Thus, it is important to consider it. Several studies have been published in this regard, most of which depend on radar sensors, ultrasound, lidar and other technologies [12, 13]. With the development of learning techniques, dependence on the CNN algorithm has emerged because it can finish the learning task to estimate the depth of obstacles and discover them, such as the model presented by Mancini et al. as the accuracy of deep learning can make a real revolution in this field [14]. Also, Parmar et al. proposed an improved model based on CNNs based on adding multiple layers that are useful for estimating the range [15]. Through this model, they were able to discover, arrange and classify obstacles at the same time. The advantage of microscopic vision over monocular is its ability to obtain 3D information directly and for the entire scene. In addition to its ability to complete the relationship between the road surface and obstacles, it depends on the engineering characteristics.

In fact, the trend is towards sequential learning due to its strong characteristics, where the features are classified based on the ease of absorption, so the features that are difficult to assimilate may be given more weight than the normal features that are commonly used. Given the deep learning of object recognition, this problem is urgent. Therefore, deep learning models are relied upon to solve the difficult computation using the sliding window model.

IV. IMPLEMENTATION STEPS

Autonomous driving technology relies on multi-purpose cameras and sensors that accurately recognize the environment as well as computer network algorithms that correctly analyze the current situation. The stereo camera installed on the windshield, the camera on the side-view mirror, the ultrasonic sensors on the front and rear bumpers, and the long-range and short-range radar detect the movement of objects. If a vehicle's computer network identifies a potentially hazardous situation, it either alerts the driver about the conditions or brakes the vehicle to prevent an accident. All cameras and sensors are intertwined like a spider web to allow the vehicle to effectively respond to danger by sharing and analyzing information. The key to this technology is to integrate data from all sensors to anticipate such situations in advance and anticipate what the driver may not expect.

When it comes to self-driving cars, the control system makes navigation decisions. The system is able to find the optimal route to its destination without hindrance through its internal map. Once the best route is determined, commands are sent to each vehicle operator to control steering and braking.

Autonomous driving is evolving in line with developments in mechanical performance, such as the accuracy of cameras, sensors, and control systems. Currently, advances based on this technology allow cars to drive while maintaining distance between them and the vehicle ahead, keeping the car in its lane, reading traffic lights through cameras, and slowing down before entering intersections or corners by attaching itself to it. Navigation Maps. In the future, it is expected that cars will be able to recognize and prepare in advance for any potential hazards on navigational routes using cloud servers.

V. CONVOLUTIONAL NEURAL NETWORK (CNN)

Layered networks have shown success due to their ability to exploit the synthetic architecture of natural data. A group of objects in a given layer creates a new element in the next layer, with hierarchical combinations. If we simulate this hierarchy as a group of layers and leave the network the task of extracting and learning the appropriate characteristics for it, we will have created what is known as a deep learning model. Hence, it can be said that deep learning networks are hierarchical networks [16].

Most people are lazy by nature, so we will definitely need machines to perform mathematical operations. The beauty of it is that this machine will do the job in a much better way than some people will do when carefully analyzing all the price dependencies in their minds. In fact, this kind of problem was

the main catalyst for the birth of machine learning. This advanced technology enables autonomous cars to distinguish between pedestrians and other vehicles and the landscapes surrounding them, and this technique is known as deep learning; It is a branch of machine learning, which revolves around processing different formats of data, such as images, video, and text through multiple levels of representation and abstraction to extract the largest possible amount of information [17].

Deep learning is a branch of artificial intelligence that seeks to replicate our capabilities to be taught and developed in machines. In other words, deep learning allows computers to receive new information, decode data, and process output, all without any human assistance. This field has huge implications for the technology of the future, including self-driving cars. People train CNNs by indoctrinating them to images previously processed using these inputs, so the algorithm continuously adjusts the weight or weight it places on each node and learns how to identify patterns and points of interest on its own. When the algorithm knows which of the nodes is the most important, it becomes more accurate; it often outperforms humans [18]. The CNN algorithm is one of the most efficient algorithms in autonomous vehicle systems due to its ability to detect, detect, distinguish and other characteristics necessary in the process of discovering and distinguishing objects and people in the vicinity of the vehicle, as shown in Fig.1.

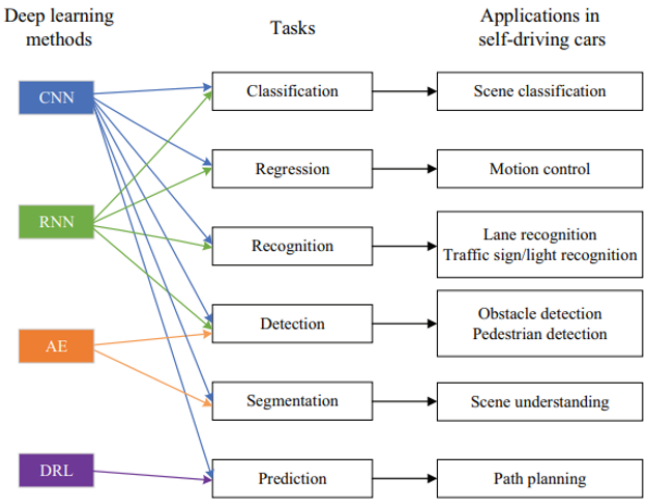

Fig. 1: Deep learning used algorithms for automated vehicles [19]

VI. Implementation Details

In this section, we discuss working methods and the conceptual framework for the proposed method of automated vehicles to be autonomous, based on the CNN algorithm.

A. Data Preprocessing

Generally, images are captured as a 20-frame set of data. Accordingly, there will be many interlocking tires. When the image is captured, it is a wide-range image that contains many unnecessary impurities. As a first step, the part near the vehicle is trimmed to be processed faster than if the entire image was taken in the processing. After cropping the image, it is enlarged, but this enlargement will lead to a difference in contrast and clarity and thus overlap the pixel values, so it must be addressed before starting the analysis.

B. Vehicles-assisted Image Sharing

Modern communication networks allow us to transfer a large amount of data at a high volume and speed, approximately (gigabits/second). Thus, it is possible to benefit from this huge amount of data through its collection and analysis. We suggest that data be collected from two cars, one behind the other, as large data will be generated at the same time, so we collect the captured images and analyze them in real time. For example, we will generate successive frames for each vehicle. We take these frames from the first car and compare them with the frames coming from the second vehicle, so they are in the form (t, t − 1, ..., t− x + 1). Thus, we will have data that can be analyzed based on the data of the first car that is located at the front.

C. Traffic Light Detection

One of the most important applications that must be considered in autonomous vehicles is traffic signals. Since we are talking in this paper about discovering objects and people and interacting with them in autonomous vehicles, traffic signals are among the static, immobile objects that must be distinguished. These signals are distinguished by analyzing and comparing the captured frames to distinguish colors and shapes, and for this, a lightweight convolutional neural network classifier (CNN) is used. As shown in Fig. 2, a diagram of autonomous vehicles using the CNN algorithm, based on data collection through cameras placed in the vehicle.

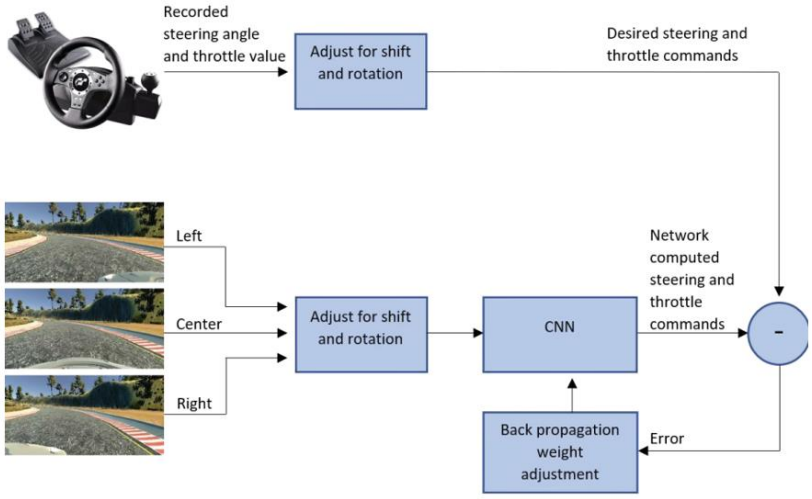

Fig. 2: Self-driving vehicles procedure using CNN algorithm [20]

VII. JUST-IN-TIME FEATURES

CB. Checkerboard (CB) highlights are acquired by connecting the ACF channels to a bunch of checkerboard channels. [21] It showed that a straightforward mix of these highlights could do very well for person on foot discovery. In view of their perception that the quantity of highlights decides the presentation (as opposed to the kind of highlight), we receive a bunch of eight basic chessboards two channels in Fig. 3. Altogether, there are 16 8 80 = 10240 CB highlights for every common zone. LDA. Privately corresponded Hoard highlights, figured by straight separation investigation (LDA), showed some unrivaled significance for identifying life forms on Hoard highlights [22]. [23] showed that computing these highlights on ACF channels prompts a critical improvement over ACF. We embrace this trademark family, with a three-channel size. Altogether, there are 16 8 40 = 5120 LDA qualities for every common fix.

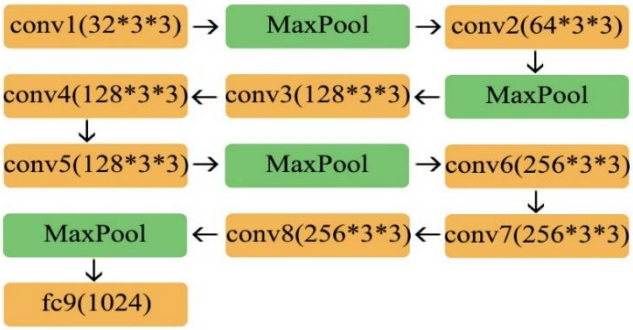

Fig. 3: CNN architecture used to extract small CNN features

VIII. EMBEDDING LARGE CNN MODELS

Large CNNs [24-26] are famous in PC vision. While, hypothetically, they could be utilized in any course stage, this makes the iterative boosting streamlining excessively computationally concentrated. It is viable, in any case, to utilize a huge CNN as the last course stage. This can, on a fundamental level, be accomplished for any CNN finder in writing. We next discuss the installing of a couple of famous techniques

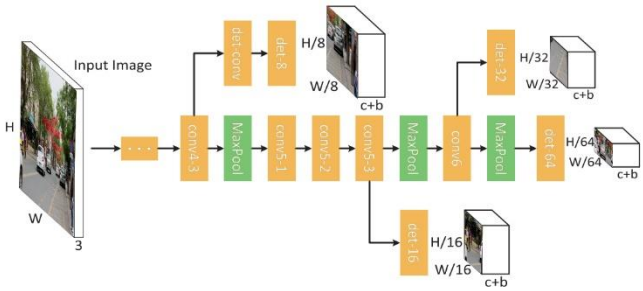

Fig. 4: MS-CNN sub-network proposal. Bold blocks are the output tensors. c, b refer to the number of classes and bounding-box coordinates

IX. EXPERIMENTAL PART

Data analysis begins with relying on a series of decision trees consisting of 2048 decision trees with depth 2 that are trained through seven consecutive initial stages of f32 decision trees; 96;128;256;512;512; 512g. After training, standard evaluation metrics are used in mAP (mean average precision) on MR (log-average miss-rate) on Caltech. Then, the obtained images are analyzed by distinguishing the dimensions of the objects in the images after cropping and enlarging them to identify pedestrians and static objects.

A. Pedestrian detection

The CityPersons dataset is relied upon, which includes a large number of images, approximately 5000 images. What distinguishes this dataset is its capture in several different cities and in different seasons and times, which gives more accuracy and credibility that simulates reality where we can rely on it.

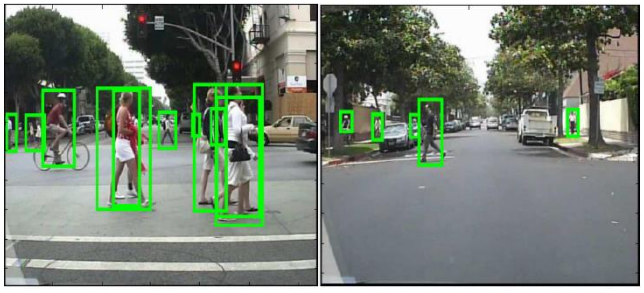

Fig. 5: Pedestrian discrimination based on CityPerson dataset

The dataset is divided into two parts (training 2975 and testing 1525) in addition to the presence of thousands of

images that contain six or seven people who can be distinguished, as well as the presence of many objects that can be distinguished. The system is trained on the training block and evaluated based on the validation set, and images are enlarged by 1.3.

B. *Object detection*

One of the most important challenges facing pedestrian detection systems is the presence of a mixture of people and moving and standing objects. Therefore, we rely on Caltech Dataset, one of the most popular datasets used for this purpose because it includes approximately 10 VGA video recordings of the streets using a car that is cruising in it. Moreover, it includes groups (training 42782 and testing 4024). It also includes many poor-quality images that can also be used in the training process in anticipation of any emergency that may encounter the vehicle on the street. We trained the system using the image scale *2 and the learning rate was 0.05, and it is gradually reduced.

X. RESULTS

In Table I , we evaluate the final model by comparing it with the results obtained in previous work TLL (MRF) [27], faster adaptation of RCNN [28], OR-CNN [29], ALFNet [30], PODE + RPN [31], Repulsion loss [32]. Our proposed method achieves remarkable success by reducing the MR value to 10.14% compared to the best results presented by [33].

TABLE I: RESULTS OF PERFORMANCE ANALYSIS IN A VALIDATION SET BASED ON CITYPERSONS DATASET

Method	TLL(MRF)	Adapted FasterRCNN	ALFNet	Repulsion Loss	PODE+RPN	OR-CNN	**Proposed**
Backbone	ResNet-50	VGG-16	VGG-16	ResNet-50	VGG-16	ResNet-50	**ResNet-50**
Scale	-	*1.3	*1	*1.3	-	*1.3	***1.3**
Reasonable	14.4	12.97	12	11.6	11.24	10.23	**10.14**

An interesting feature of CompACT is that the detector sequence can be thought of as an attention mechanism that assigns calculations to image positions based on the complexity of the detection. Fig. 6 illustrates this feature and summarizes the calculations spent on each Caltech test image. The time complexity varies considerably with the content of the image. The image on the left shows a simple scene with few objects and no pedestrians. Since the simple and powerful features are enough to get rid of all windows, the complexity of image processing is low. The image at the far right shows a more complex scene with varying proportions of detail (buildings in the foreground and buildings in the background) and many pedestrians. In this case, many areas of the image propagate to the final stage of the chain, and the complexity is high. In this sense, CompACT behaves more like the human visual system. For example, a human commentator quickly realized that the left image does not contain pedestrians, but it takes longer to count the number of pedestrians in the rightmost image. This behavior is very different from channel-level object detectors [34,35] and CNN-based object detectors [36,37], whose time complexity is almost independent of the image content. From an application point of view, this has both advantages and disadvantages. For example, in Fig. 6, the processing time range is from 0.05 to 2.0 s per frame, and the variance is approximately 0.04 s, which may cause problems for applications that require continuous processing time. For these, a channel or CNN detector may be a better solution. On the other hand, there are benefits to customizing your account according to your needs, especially if you have power issues. For example, running a Titan X GPU card 24 hours a day requires 6 kWh per day, which is six times the energy consumption of a refrigerator. This alone can prevent applications from being deployed in-home monitoring, edge devices and drones.

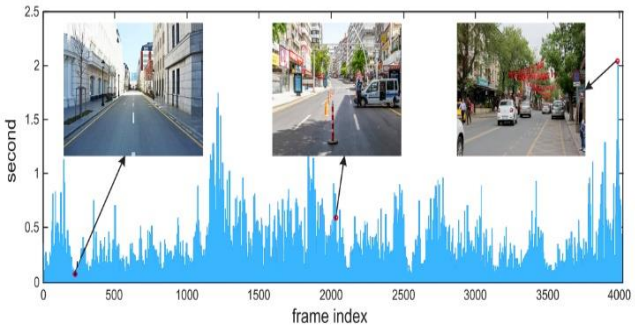

Fig. 6: Processing time spent per Caltech test image.

XI. CONCLUSION

Autonomous vehicles are one of the most critical current applications towards automating operations, and the most important of which is the transportation sector. One of the most significant challenges facing researchers in this field is the ability to discover and distinguish objects, whether they are moving or stationary. Since the significance of the detected objects is arranged according to precedence, for example, pedestrians have the highest priority and then traffic lights. In this paper, a model for self-driving vehicles is proposed that helps in detecting objects and pedestrians in the vicinity of the vehicle, depending on the techniques of complex and reinforcement learning, and for this matter, it relies on the CNN algorithm. The strengths of this model and the reasons for its potential success are discussed. The expected results of implementing this model were also reviewed.

In future work, the focus will be on the response time to implement the discovery and distinguish things in the least possible time to increase the efficiency of the system to give a greater safety ratio and a better ability to deal with different conditions.

REFERENCES

[1] M. L. MUTAR, M. A. BURHANUDDIN, A. S. HAMEED, N. YUSOF, M. F. ALRIFAIE, and A. A. MOHAMMED, "Multi-objectives ant colony system for solving multi-objectives capacitated vehicle routing problem," J. Theor. Appl. Inf. Technol., vol. 98, no. 24, pp. 4014–4027, 2020.

[2] W. Ouyang, H. Zhou, H. Li, Q. Li, J. Yan, and X. Wang, "Jointly learning deep features, deformable parts, occlusion and classification for pedestrian detection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 8, pp. 1874–1887, Apr. 2018.

[3] P. Dollar, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 8, pp. 1532–1545, Aug. 2014.

[4] W. Ouyang, H. Zhou, H. Li, Q. Li, J. Yan, and X. Wang, "Jointly learning deep features, deformable parts, occlusion and classification for pedestrian detection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 8, pp. 1874–1887, Apr. 2018.

[5] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum pointnets for 3d object detection from RGB-D data," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2018, pp. 918–927.

[6] Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan, "Scale-Aware Fast {R-CNN} for Pedestrian Detection," IEEE Trans. Multimedia, vol. 20, no. 4, pp. 985–996, 2018, https://doi.org/10.1109/TMM.2017.2759508.

[7] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Strengthening the effectiveness of pedestrian detection with spatially pooled features," in Proc. Eur. Conf. Comput. Vis., 2014, pp. 546–561

[8] S. Zhang, R. Benenson, and B. Schiele, "Filtered channel features for pedestrian detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 1751–1760.

[9] X. Shao et al., "Pedestrian Detection Algorithm based on Improved Faster RCNN," IEEE Adv. Inf. Technol. Electron. Autom. Control Conf., vol. 2021, pp. 1368–1372, 2021, doi: 10.1109/IAEAC50856.2021.9390882.

[10] S. Zhang, ; J. Yang, .; and B. Schiele, . 2018. Occluded pedestrian detection through guided attention in cnns. In CVPR.

[11] C. Zhou, and J. Yuan,. Bi-box regression for pedestrian detection and occlusion estimation. In ECC, 2018 .

[12] Liu, H.; Sun, F.; Zhang, X. Robotic material perception using active multimodal fusion. IEEE Trans. Indust. Electron. 2019, 66, 9878–9886.

[13] ZX. hang, M. Zhou, P. Qiu, Y. Huang, Li, J. Radar and vision fusion for the real-time obstacle detection and identification. Indust. Robot 2019, 46, 391–395.

[14] M. Mancini, M.; Costante, G.; Valigi, P.; Ciarfuglia, T.A. J-MOD2 : Joint monocular obstacle detection and depth estimation. IEEE Robot. Autom. Lett. 2018, 3, 1490–1497.

[15] H.M. Chen, vision-based obstacle detection and avoidance for a multicopter. IEEE Access 2019, 7, 167869–167883.

[16] K. Lu, J. Li, X. An, H. He and X. Hu, "Generalized Haar Filter based CNN for Object Detection in Traffic Scenes" in 13th IEEE Conference on Automation Science and Engineering (CASE) 2017, 978-1-5090-6781-7/17/$31.00.

[17] S. Jung, U. Lee, J. Jung, D.Hyunchul Shim, "Real-time Traffic Sign Recognition System with Deep Convolutional Neural Network" in 13th International Conference on Ubiquitous Robots and Ambient Intelligence (URAl) 2016, 978-1-5090-0821- 6/16/$31.00.

[18] K. Behrendt, L. Novak, R. Botros, "A Deep Learning Approach to Traffic Lights: Detection, Tracking, and Classification" in IEEE International Conference on Robotics and Automation (ICRA) Singapore 2017, 978-1-5090-4633-1/17/$31.00

[19] Ni J, Y. Chen , Y. Chen , J. Zhu , D. Ali , Cao W. A survey on theories and applications for self-driving cars based on deep learning methods. Appl Sci 2020;10:1–29. https://doi.org/10.3390/APP10082749.

[20] D. Egio J, L.M. Bergasa , E. Romera , Gómez Huélamo C, Araluce J, Barea R. Self-driving a Car in Simulation Through a CNN. Adv Intell Syst Comput 2019;855:31–43. https://doi.org/10.1007/978-3-319-99885-5_3.

[21] S. Zhang, R. Benenson, and B. Schiele, "Filtered channel features for pedestrian detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 1751–1760.

[22] B. Hariharan, J. Malik, and D. Ramanan, "Discriminative decorrelation for clustering and classification," in Proc. Eur. Conf. Comput. Vis., 2012, pp. 459–472.

[23] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 4, pp. 743–761, Apr. 2012

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–77

[25] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in Proc. 32nd Int. Conf. Mach. Learn., 2015, pp. 448–456.

[26] Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," ICLR, 2015, http://arxiv.org/abs/1409.1556.

[27] T. Song,. L. Sun, Xie, D.; Sun, H.; and Pu, S. Smallscale pedestrian detection based on topological line localization and temporal feature aggregation. In ECCV, 2018. .

[28] S. Zhang, R. Benenson, and B. Schiele, . Citypersons: A diverse dataset for pedestrian detection. In CVPR, 2017 .

[29] S. Zhang, Wen, L.; Bian, X.; Lei, Z.; and Li, S. Z. Occlusion-aware R-CNN: detecting pedestrians in a crowd. In ECCV, 2018.

[30] W. Liu, S. Liao, Hu, W.; Liang, X.; and Chen, X. Learning efficient single-stage pedestrian detectors by asymptotic localization fitting. In ECCV, 2018.

[31] C. Zhou, and Yuan, J. Bi-box regression for pedestrian detection and occlusion estimation. In ECCV, 2018.

[32] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and Shen, C. Repulsion loss: Detecting pedestrians in a crowd. In CVPR, 2018.

[33] C. Chi ,S. Zhang Xing J, Lei Z, Li SZ, Zou X. Relational learning for joint head and human detection. AAAI 2020 - 34th AAAI Conf Artif Intell 2020:10647–54. https://doi.org/10.1609/aaai.v34i07.6691.

[34] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Strengthening the effectiveness of pedestrian detection with spatially pooled features," in Proc. Eur. Conf. Comput. Vis., 2014, pp. 546–561.

[35] P. Dollar, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 8, pp. 1532–1545, Aug. 2014

[36] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multiscale deep convolutional neural network for fast object detection," in Proc. Eur. Conf. Comput. Vis., 2016, pp. 354–370

[37] L. Zhang, L. Lin, X. Liang, and K. He, "Is faster R-CNN doing well for pedestrian detection?" in Proc. Eur. Conf. Comput. Vis., 2016, pp. 443–457.