Improving Traffic Surveillance with Deep Learning Powered Vehicle Detection, Identification, and Recognition

Priyanka Patel[1[0000-1111-2222-3333]], Rinkal Mav[2[1111-2222-3333-4444]], Pratham Mehta[3[2222-3333-4444-5555]], Kamal Mer[4[3333-4444-5555-6666]], Jeel Kanani[5[4444-5555-6666-7777]]

[1,2,3,4,5] Smt. Kundaben Dinsha Patel Department of Information Technology, Chandubhai S. Patel Institute of Technology, Faculty of Technology & Engineering, Charotar University of Science and Technology (CHARUSAT), Changa-388421, Gujarat, India.
priyankapatel.it@charusat.ac.in

Abstract. As the volume of vehicles on our roads continues to surge, accurate detection and counting of vehicles have become critical for effective traffic management. Identifying vehicles precisely is challenging due to the wide range of sizes, shapes, and external factors influencing computer vision. To overcome these challenges, here propose a vehicle detection strategy based on the YOLOv5 algorithm. YOLOv5 is an advanced object detection algorithm leveraging convolutional neural networks (CNNs) for high-precision, high-speed detection in images and videos. Our strategy harnesses YOLOv5's capabilities, optimizing it for both speed and accuracy. Comprising convolutional layers, pooling layers, and fully-connected layers, YOLOv5 collaboratively detects and identifies vehicles in images or video frames. Extensive training on a diverse dataset empowers the algorithm to recognize vehicles with exceptional precision. An empirical study evaluated YOLOv5's performance across diverse vehicle types and environmental conditions. Results unequivocally demonstrated substantial improvements in vehicle detection speed and precision. Even under challenging scenarios, the algorithm consistently achieved real-time identification and enumeration of vehicles.

Keywords: Computer Vision, Deep Learning, Traffic surveillance system, convolutional neural networks, YOLOV5, Detection.

1 INTRODUCTION

Undoubtedly, the quality of urban living faces a severe threat owing to the rapid global expansion of urbanization and the escalating issue of traffic congestion in many nations across the world [1]. The perils associated with travel and exposure to traffic-related hazards increase significantly as motorization rates soar and road networks expand. This is primarily because the number of registered vehicles consistently outpaces population growth, resulting in more roads being constructed. Road traffic accidents, in particular, stand as one of the leading causes of fatalities, debilitating injuries, and hospitalizations in contemporary society, imposing a substantial socioeconomic burden worldwide. According to the "Road Accidents in India 2020" report [2], a staggering 3,66,138 road accidents occurred in 2020 alone, resulting in 1,31,714 fatalities and 3,48,279 hospitalizations. The emergence of artificial intelligence has ushered in the era of autonomous transportation, facilitating the development of the Intelligent Transportation System (ITS). This technological advancement significantly enhances both driver and passenger safety and comfort. ITS leverages cutting-edge technologies such as information and communication technologies, intelligent control systems, and electronic sensors [3], offering an array of intelligent services that encompass emergency management, dynamic traffic control, and real-time passenger alerts for public transportation schedules. In this context, vehicle detection emerges as a pivotal task critical to the success of these applications. One vital application of vehicle detection systems is in the realm of traffic camera systems [3,4]. These systems primarily comprise automated surveillance digital cameras designed to capture images of passing vehicles and other moving objects. High-resolution static images obtained through this method provide law enforcement agencies and security departments with essential details about vehicle identification, passage time, route, and other pertinent information. Previously, vast archives of these images demanded manual processing, a labor-intensive and yield-limited endeavor. However, due to the rapid evolution of computer vision technology, the latest vehicle license plate recognition software is being increasingly adopted within this domain, yielding impressive results.

Nevertheless, challenges persist in this domain, as there are instances where the recognition model fails to identify a vehicle's license number due to factors such as cloned plates, missing plates, or plates that are unreadable. This underscores the growing need for automatic vehicle identification and recognition systems in traffic monitoring [5, 6]. Consequently, the development of an effective and reliable detection system becomes an imperative endeavor. However, the inefficiency of vehicle detection can lead to a decline in security levels. In this paper, the author proposes the utilization of the YOLOv5 enabled vehicle detection model, which is considered a superior method for vehicle detection due to its exceptional balance between speed and detection accuracy. The author conducts experiment on the model trained using YOLOv5 and presents the findings. Notably, there is no further comparison made with other objection models in this study.

2 WORK DONE RELATED TO THE TOPIC

In classical vehicle detection techniques, the conventional approach often involves manual feature extraction, with the objective of identifying moving vehicles within video sequences. This process entails extracting pertinent features and subsequently classifying them using a designated classifier [7,8]. Various methods have been employed for the detection of vehicles, and among them are: (1) The Background Update Approach: This method relies on the concept of a weighted average to continually update the background. Its application often has implications for the precision of target detection and the completeness of target extraction [9, 12]. (2) The Frame Difference Method: This approach involves calculating the difference between successive frames in order to achieve target extraction. Factors such as the vehicle's speed and the time gap between consecutive frames significantly influence the effectiveness of this method [10, 13]. (3) The Optical Flow Method: The optical flow method operates at the pixel level and is utilized for density estimation [11,14]. However, with the advent of deep learning and modern object detection frameworks, newer and more accurate methods for vehicle detection have been developed, which have shown significant improvements in accuracy and speed.

Regions with Convectional Neutral Network(R-CNN): R-CNN variety of operating methods, which is based on the candidate region's two stages. The two-stage method's initial phase involves creating a candidate box in a certain manner. The candidate box's contents are then categorised and discriminated upon, and its position is adjusted. The reason it has two stages—generating candidate regions and identifying them—is why it is known as the two-stage approach. Initially, the primary deep learning approach used for target detection was the Convolutional Neural Network (CNN). One of the first methods to employ CNNs for object detection was R-CNN [9], which was introduced by Ross Girshick et al. The term R-CNN stands for "Regions with CNN features," with the "R" denoting region. In this technique, the target objects were first identified by extracting features using CNNs in regions of interest that had been exhaustively searched. CNNs had already demonstrated their effectiveness in whole-image classification tasks, and this was leveraged for more complex tasks like object detection that required advanced feature extraction beyond simple image classification.

SPP-net: In 2014, Kaiming Heetal developed SPP-net, which introduced the spatial pyramid pooling method that greatly improved the performance of R-CNN. This new method addressed the challenge of time-consuming feature extraction by applying the CNN feature calculation to each candidate region in R-CNN, resulting in a significant increase in detection speed. It also resolved the issue of target shape distortion caused by scaling the detection image to match the input size of CNN. With the use of SPP-net, only one CNN calculation is required for each test image, and feature extraction is performed on the entire inspection image[15, 27,29].

Fast CNN: It was a revised version of R-CNN that used the SPP-NET method previously introduced by RBG. The full ROI feature map was used to directly extract features, resulting in improved efficiency. Faster R-CNN [16, 17] was an even more advanced version that was optimized for faster detection speed.

YOLO (You Only Look Once): Yolo family of object detection methods, a distinctive characteristic is the one-stage approach, which diverges from the traditional two-stage method. In the one-stage approach, candidate boxes are not generated beforehand; instead, they are instantly predicted and classified across the entire image. Notable examples of single-stage approaches include SSD [24] and YOLO [18,28]. The inception of the single-stage detector can be traced back to Joseph's work in 2015. YOLO, in particular, made a significant impact by surpassing RCNN in terms of both accuracy and speed, achieving an impressive 45 frames per second with a quicker version clocking in at 155 frames per second. YOLO marked the beginning of single-stage detection algorithms and played a pivotal role in facilitating the practical implementation of target detection in industrial settings. In this context, a novel concept emerged within YOLO, where CNNs were applied to process the majority of images directly, employing direct regression of output characteristics to predict target categories and bounding boxes. The evolution of the YOLO family continued in 2017 with the release of YOLOv2, which garnered an honorable mention at CVPR. This architecture built upon the foundation of YOLO by introducing a range of incremental improvements, including the incorporation of BatchNorm, higher resolution, and anchor boxes. Subsequently, in 2018, YOLOv3 was introduced, enhancing earlier models by introducing an objectness score into bounding box prediction to improve performance on smaller objects. YOLOv3 also increased connections to the backbone network layers and made predictions at multiple levels of granularity.

The release of YOLOv3, its original creator, Joseph Redmon, departed from computer vision research. The "YOLO family" saw its fourth publication in April 2020 with the introduction of YOLOv4, authored by Alexey Bochkovskiy, marking the first instance where Joseph Redmon was not the author. YOLOv4 brought numerous enhancements, including augmentations, mish activation, improved feature aggregation, and various other upgrades. In June 2020, Glenn Jocher released the fifth installment in the "YOLO family," known as YOLOv5. This version was aimed at expediting model training and reducing experimentation costs. Consequently, the improved YOLOv5 version gained prominence in research projects, including those that involved gathering extensive datasets encompassing diverse vehicle types.

However, YOLOv5 exhibited limitations when confronted with adverse weather conditions, such as foggy or inclement weather, where visibility is severely compromised. In such scenarios, the algorithm's performance deteriorated. To address this challenge, an extended version of YOLOv5, known as YOLOv5-Fog, was developed for object detection under adverse weather conditions. YOLOv5-Fog demonstrated an increased accuracy rate of 73.4%, a noteworthy improvement of 5.4% over the standard YOLOv5. This enhancement was achieved through the incorporation of YOLOv5 features and the integration of SwinFocus, the decoupled head model, along with the replacement of conventional non-maximum suppression methods with Soft-NMS. These adaptations collectively contributed to heightened accuracy and improved results in adverse weather conditions [30].

3 DATASETS

In Table 1, an illustrative dataset is presented, comprising images of vehicles, specifically cars, that were acquired during journeys through diverse neighborhoods in Bangalore and Hyderabad. The dataset encompasses images of varying resolutions, with the majority being at a resolution of 1080p. However, a subset of images exists at a resolution of 720p, while some exhibit different resolutions [20]. To facilitate the development and evaluation of the model, the dataset has been meticulously divided into three distinct sets: training, validation, and test sets. Each of these sets plays a crucial role in the model's development and the assessment of its performance.

Table 1 Summary of collected dataset images.

Database Division	**Total Images**
Total Images (collected from Video)	46,588
Trained	31,569
Validated	10,225
Tested	4,794

Label Statistics:

Here the Label Statistics means that for a given dataset the each and every data is labelled and by using the labelling facility we can statistical analyse the data in the form of graph. By taking the above dataset we can make histogram of pixel distribution. (1) Pixel counts for each label in Y-axis. (2) Label names in X-axis.

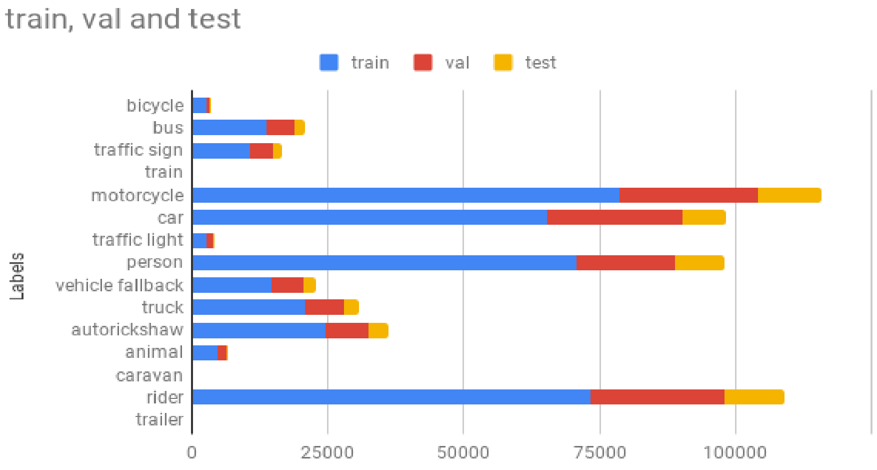

Fig. 1 Histogram of pixel Distribution

4 PROPOSED METHOD:

Yolo, as a state-of-the-art real-time object detection model that has undergone significant advancements through its various iterations, from Yolov1 to Yolov4, with Yolov5 being the latest milestone. Over the years, its performance has consistently elevated, achieving remarkable results on benchmark object detection datasets like Microsoft COCO (common objects in context) [21] and Pascal VOC (visual object classes) [22]. Figure 2 provides a visual representation of the Yolov5 network architecture [23].

The evolution of the Yolo family of models has played a pivotal role in advancing the realm of object detection, culminating in more resilient and accurate techniques within the field of computer vision. The Yolov5 architecture comprises three fundamental components, each contributing uniquely to its functionality: ***CSPDarknet:*** Serving as the backbone of the model, CSPDarknet undertakes the initial feature extraction from the input data. This module plays a critical role in capturing essential image features. ***PANet:*** Positioned as the neck of the architecture, the PANet module is responsible for the fusion of extracted features with other relevant features. This fusion process enhances the model's capability to comprehend intricate patterns and relationships within the data. ***Yolo Layer:*** Acting as the head of the architecture, the Yolo Layer assumes responsibility for the detection process itself. It receives the processed data from earlier stages and produces the final detection results, including the identification and localization of objects within the input.

It is imperative to highlight that the input data first undergoes feature extraction through the CSPDarknet component. Subsequently, these features are forwarded to the PANet module, where they are combined with other learned features, further enriching the model's understanding of the data. Finally, the Yolo Layer generates the ultimate detection results, making Yolov5 a comprehensive and effective object detection tool within the realm of computer vision.

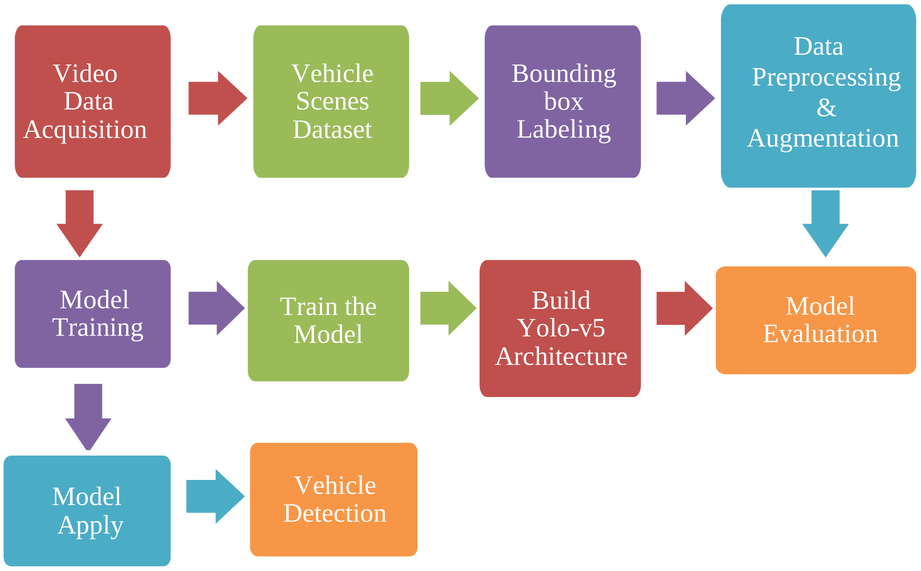

Fig. 2 Proposed Architecture of detecting vehicle

The decision to adopt Yolov5 as the initial learning model is underpinned by three key factors, each contributing significantly to its suitability for the task at hand.

Firstly, Yolov5's utilization of the cross-stage partial network (CSP-Net) [24], seamlessly integrated into Darknet's CSPDarknet backbone, stands out as a pivotal factor.

The incorporation of CSP-Net within CSPDarknet addresses the issue of repeated gradient information in large-scale backbones, resulting in several advantageous outcomes. It notably reduces the number of model parameters and floating-point operations per second (FLOPS), leading to a more streamlined model that boasts enhanced inference speed and accuracy. Moreover, this integration facilitates model size compression without compromising performance. In the context of detecting moving vehicles, where the swiftness and precision of detection are paramount, the model's compactness plays a crucial role. It ensures efficient inference on edge devices equipped with limited resources.

The second pivotal factor is Yolov5's adoption of the path aggregation network (PANet) [4, 25] within its architecture. PANet enhances information flow and optimizes overall detection performance. It achieves this by facilitating the transmission of low-level features, complemented by a feature pyramid network (FPN) structure that strengthens the bottom-up path. The synergy of these techniques results in improved feature propagation, ensuring the efficient utilization of all features to enhance detection performance. The novel FPN structure within the PANet module optimizes information flow and effectively extracts features across various scales, thereby bolstering the model's overall effectiveness. Adaptive feature pooling, connecting the feature grid to all feature levels, expedites the dissemination of vital information from each feature level to subsequent subnetworks. Furthermore, the PANet module enhances the utilization of precise localization signals in the lower layers, resulting in superior object location accuracy.

Lastly, the Yolov5 model incorporates the Yolo Layer as its head, enabling multi-scale prediction. This feature generates feature maps in three distinct sizes (18x18, 36x36, and 72x72), a critical capability for object detection tasks where the accurate identification and localization of objects of varying sizes within a scene are essential.

Yolov5's adoption of these techniques results in a model that excels in detection performance, combining enhanced accuracy and speed. These attributes render it highly suitable for real-world applications, particularly in the context of detecting moving vehicles where efficiency and precision are paramount.

5 Results - Detected Vehicles from Videos

In the YOLOv5 model, trained vehicles undergo object detection. The process commences with the selection of a single frame or image from the dataset. This chosen image is subsequently utilized to train the YOLOv5 model. Through this training process, the YOLOv5 model effectively identifies and categorizes the specific type of vehicle present in the image.

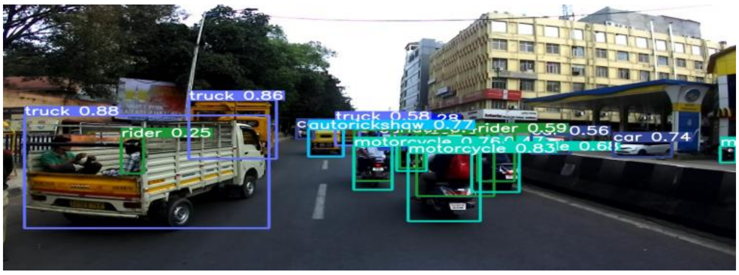

Fig. 3 Vehicles detected from the frame/image

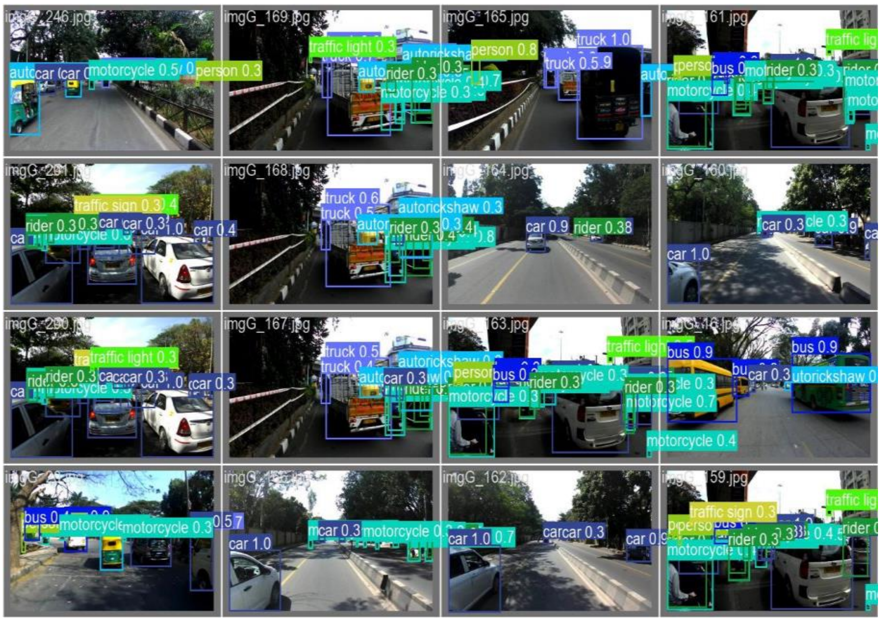

Fig. 4 Multiple Vehicle Detected and recognized in Different Frames

The results indicate that, based on the substantial volume of training data utilized, the YOLO-V5 model achieves outstanding levels of accuracy and precision. Specifically, it attains a mean average precision (mAP) of 49.9%, underlining its proficiency in object detection and classification.

CONCLUSION AND FUTURE WORK

Convolutional neural networks have proven highly effective in significantly improving object detection performance. The primary objective of this study was to harness the power of YOLOv5 to enhance the vehicle detection system for recorded images. These investigations were conducted using Google Colab, an online platform, yielding the following outcomes: A mean average precision of 0.499, with a threshold set at 0.5, was achieved. While the precision attained is commendable, there is room for further enhancing accuracy. Future research endeavors will prioritize bolstering the model's robustness and adaptability, including expanding the range of object classes, such as incorporating categories for emergency vehicles like ambulances. Additionally, efforts will be directed towards augmenting the dataset with a wider array of real-world image scenarios, while also exploring applications involving vehicle count tracking, offering potential benefits across diverse domains.

References

1. Abrougui, Alia, and Mohamed Hayouni. "Convolutional Neural Network for Vehicle Detection in A Captured Image." 2022 International Wireless Communications and Mobile Computing (IWCMC). IEEE, 2022.
2. Patel, Priyanka, and Amit Nayak. "Predictive Convolutional Long Short-Term Memory Network for Detecting Anomalies in Smart Surveillance." Reliability: Theory & Applications 17.3 (69) (2022): 139-161.
3. L. Zhu, F. R. Yu, Y. Wang, B. Ning and T. Tang, " Big Data Analytics in Intelligent Transportation Systems: A Survey," in IEEE Transactions on Intelligent Transportation Systems, Jan. 2019, pp. 383-398, .
4. Patel, Priyanka, and Amit Nayak. "Predictive Convolutional Long Short-Term Memory Network for Detecting Anomalies in Smart Surveillance." Reliability: Theory & Applications 17.3 (69) (2022): 139-161.
5. Patel, Priyanka, and Amit Thakkar. "Machine Learning Techniques To Detect Anomalies In Surveillance Videos." IJRAR-International Journal of Research and Analytical Reviews (IJRAR) 5.4 (2018): 204-207.
6. Zheng, X.; Chen, F.; Lou, L.; Cheng, P.; Huang, Y. Real-Time Detection of Full-Scale Forest Fire Smoke Based on DeepConvolution Neural Network. Remote Sens. 2022, 14, 536.
7. Zhao, H.; Li, Z.; Zhang, T. Attention Based Single Shot Multibox Detector. J. Electron. Inf. Technol. 2021, 43, 2096–2104.
8. Patel, Priyanka, and Amit Thakkar. The upsurge of deep learning for computer vision applications." International Journal of Electrical and Computer Engineering 10.1 (2020): 538.
9. Lee, D.S. Effective Gaussian mixture learning for video background subtraction. IEEE Trans. Pattern Anal. Mach. Intell. 2005, 27,827–832.
10. Deng, G.; Guo, K. Self-Adaptive Background Modeling Research Based on Change Detection and Area Training. In Proceedingsof the IEEE Workshop on Electronics, Computer and Applications (IWECA), Ottawa, ON, Canada, 8–9 May 2014; Volume 2,pp. 59–62.
11. Muyun, W.; Guoce, H.; Xinyu, D. A New Interframe Difference Algorithm for Moving Target Detection. In Proceedings of the2010 3rd International Congress on Image and Signal Processing, Yantai, China, 16–18 October 2010; pp. 285–289.
12. Zhang, H.; Zhang, H. A Moving Target Detection Algorithm Based on Dynamic Scenes. In Proceedings of the 8th InternationalConference on Computer Science and Education (ICCSE), Sri Lanka Inst Informat Technol, Colombo, Sri Lanka, 26–28 April2013; pp. 995–998.
13. Barnich, O.; Van Droogenbroeck, M. ViBe: A Universal Background Subtraction Algorithm for Video Sequences. IEEE Trans.Image Process. 2011, 20, 1709–1724
14. Fang, Y.; Dai, B. An Improved Moving Target Detecting and Tracking Based On Optical Flow Technique and Kalman Filter. InProceedings of the 4th International Conference on Computer Science and Education, Nanning, China, 25–28 July 2008; pp.1197–1202.

15. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation.In Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28June 2014; pp. 580–587.
16. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. InProceedings of the 13th European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014; pp. 346–361
17. Girshick, R. Fast r-cnn. In Proceedings of the Tenth IEEE International Conference on Computer Vision, Beijing, China, 17–20 October 2005; pp. 1440–1448.
18. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings ofthe 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 27–30 Jun 2016; pp. 779–788.
19. Nelson, J. "Your Comprehensive Guide to the YOLO Family of Models." blog. roboflow. com (2022).
20. Varma, Girish, et al. "IDD: A dataset for exploring problems of autonomous navigation in unconstrained environments." 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2019.
21. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the 13th European Conference on Computer Cision (ECCV 2014), Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
22. Everingham, M.; Eslami, S.A.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes challenge: A retrospective. Int. J. Comput. Vis. 2015, 111, 98–136, doi:10.1007/s11263-014-0733-5.
23. Xu, Renjie, et al. "A forest fire detection system based on ensemble learning." Forests 12.2 (2021): 217.
24. Wang, C.Y.; Mark Liao, H.Y.; Wu, Y.H.; Chen, P.Y.; Hsieh, J.W.; Yeh, I.H. CSPNet: A new backbone that can enhance learning capability of cnn. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2020), Washington, DC, USA, 14–19 June 2020; pp. 390–391.
25. Wang, K.; Liew, J.H.; Zou, Y.; Zhou, D.; Feng, J. Panet: Few-shot image semantic segmentation with prototype alignment. In Proceedings of the IEEE International Conference on Computer Vision (ICCV 2019), Seoul, Korea, 20–26 October 2019; pp. 9197– 9206.
26. Patel, Bhunesh, Neel Ray, and Priyanka Patel. "Motion based Object Tracking." International Journal of Electronics, Electrical and Computational System 7.4 (2018): 581-588.
27. Patel, Priyanka P., and Amit R. Thakkar. "A Journey from Neural Networks to Deep Networks: Comprehensive Understanding for Deep Learning." Neural Networks for Natural Language Processing. IGI Global, 2020. 31-62.
28. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. arXiv 2018, arXiv: 1804.02767.
29. Patel, P., and A. Ganatra. "Investigate age invariant face recognition using PCA, LBP, Walsh Hadamard transform with neural network." International Conference on Signal and Speech Processing (ICSSP-14). 2014.
30. Meng, X.; Liu, Y.; Fan, L.; Fan, J. YOLOv5s-Fog: An Improved Model Based on YOLOv5s for Object Detection in Foggy Weather Scenarios. Sensors 2023, 23, 5321- 03/06/2023. https://doi.org/10.3390/s23115321