Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/50/20241583

Vehicle collaborative detection based on YOLOv5

Zikun Gong[1,4], Kezhen Guo[2], Danyao Yang[3]

[1]Tongda College, Nanjing University of Posts and Telecommunications, Yangzhou, 225127, China
[2]Taihu College, Wuxi, 214063, China
[3]Beijing University of Posts and Telecommunications, Beijing, 100876, China

[4]Corresponding author: milin@ldy.edu.rs

Abstract. Object detection has always been a popular research issue in the computer vision community. It has often been used in many downstream fields in recent years, especially vehicle and pedestrian detection in autonomous driving tasks. To improve the detection accuracy of vehicles or pedestrians, significant progress has been made in previous works. However, few of them specifically study the detection under extreme environments. Considering the need for real-time detection in research on vehicle and pedestrian detection, based on the detailed investigation of existing researches on vehicle and pedestrian detection as well as general object detection algorithms, we study the structure and framework of YOLOv5 algorithm and discuss the advantages and improvements. Then, we collect the dataset of vehicle detection based on YOLOv5 and conduct extensive experiments to analyze the detection performance. Experimental results show that our method can detect vehicles and pedestrians in different scenarios, and we hope to provide some new insights for the future development of this research field.

Keywords: Vehicle detection, pedestrian detection, YOLOv5, deep learning

1. Introduction

In recent years, object detection has received extensive attention from both academia and industry, which can greatly reduce the manpower and the error rate of recognition. Thanks to rapid development of artificial intelligence, the accuracy and speed of modern object detection technologies have made significant breakthrough, which have been widely used in various fields [1], such as intelligent security monitoring, intelligent transportation, smart homes, healthcare, and especially the autonomous driving.

As the basic technology of autonomous driving, object detection is adopted to realize the perception of surrounding high-latitude environmental information, such as roads, vehicles, pedestrians, traffic signs, etc. In recent years, great efforts have been made to improve the detection performance of vehicles and pedestrians. To avoid the impact of the environment on the test results, Zhang et al. proposed an improved algorithm for detecting people and vehicles in haze weathe based on YOLOv5, which introduces the hybrid attention mechanism and combines more powerful loss function to reduce the interference of irrelevant information in the background [2]. Extensive experiments of existing algorithms comparison, attention mechanisms ablation show this method can enhance the feature representation ability and greatly reduce the impact of haze weather on vehicle and pedestrian detection.

DOI: 10.54254/2755-2721/50/20241583

Jiang et al. change the convolutional modules with the ghost module in the backbone network, and further refine the calculation of the aspect ratio distance in the bounding box regression loss function with the calculation of the width height difference of the bounding box to accelerate network convergence [3]. This helps broaden the application scenarios of detection algorithms while maintain the high accuracy of the original algorithm. Though the aforementioned works contribute to the vehicle or pedestrian detection, we argue most of them barely study the detection under extreme environments, such as such as wind and sand weather with low visibility, rainstorm weather and other different weather, scenes with dense pedestrian flow and vehicles running in a staggered manner.

Considering the need for real-time detection in research on vehicle and pedestrian detection, this article chooses to use the YOLO series of algorithms that have performed better in object detection in recent years. The YOLOv5 algorithm requires less memory compared to other algorithms, and the weight file size of YOLOv5 is only 1/9 of YOLOv4. The training time is also relatively short in other YOLO series algorithms. In this work, based on the detailed investigation of existing researches on vehicle and pedestrian detection as well as general object detection algorithms, we study the structure and framework of YOLOv5 algorithm and discuss the advantages and improvement. Then, we collect the dataset of vehicle detection based on YOLOv5 and conduct extensive experiments to analyze the detection performance.

2. Model Construction

2.1. Experimental data acquisition

2.1.1. Annotation. We label the collected original training images with the tool of labelimg, which can get the border position and category via drawing tight boxes that covering the whole targeted objects. Put the dataset (images) into the newly created folder and create a separate folder for saving the generated data (paper_data). Next, the objects to be detected are boxed out and the categories of that detected items are entered. Finally, the annotation in .xml file format is obtained.

2.1.2. Training dataset. The first step is to transform the format of the dataset: create ‘images’ in paper_data (store .jpg image pictures), ‘Annotations’(Store .xml files), ‘ImageSets’(Create subfolder "Main"). In this paper, ImageSets are manipulated so that the following four txt documents are generated in the ImageSets/Main folder, with the purpose of partitioning the data into training sets, validation sets and so on. At the end of the data segmentation, the .xml data (image annotation) in the Annotation is parsed into .txt format. The five types of label information "class, x_center, y_center, width, height" are extracted, and three .txt files are created as the path guide for the training set, test set and validation set. Successfully generated "labels" file, which contains the labeling information of the target in .txt format, test.txt, train.txt, val.txt is the path guide. Create my.yaml in the YOLOv5-5.0/data folder, enter the path to the dataset, the number of categories in the dataset, the label and call my.yaml in train.py and provision the parameters such as epochs, batch-size, img-size and other parameters for training.

2.2. Revisiting YOLOv5

2.2.1. Feature extraction. The first key module of YOLOv5 is the feature extraction module, whose basic pipeline is summarized as following. (1) Convolution operation. Input the original image 608*608*3 into the focus structure, split into 304*304*12 characterization charts, pixel values of these small images * pixel values of 32 (3*3) sized convolution kernels, enter the results into a new pixel map and set the color shades based on the size of the new pixel values. (2) Pooling: divide the new pixel map into smaller pixel maps, take the largest pixel value in each smaller pixel map and enter it into the new pixel map, so that the new pixel map retains the most characteristic features of the original image. (3) Flattening process: The new pixel map we obtained is superimposed and processed into a one-dimensional data bar, whose entry into the fully connected hidden layer finally produces the output [4].

DOI: 10.54254/2755-2721/50/20241583

2.2.2. Edge box Realization. We define edge boxes for cars and pedestrians: converting voc's format data to YOLO's format data (x_center, y_center represents the center position of the border, width, height represents the width and height of the border respectively, through these four data can determine the edge of the box.) Next, we draw the edge boxes, and these edge boxes and the corresponding categorizations will be read in the subsequent process [5].

2.2.3. Anchor frame implementation. IoU is used to calculate the similarity between two frames,0 means no overlap,1 means overlap. A and B represent two sets. In order to prevent a lot of overlapping of anchor frames, we therefore divide the image into n*m chunks, each of which is an anchor frame. Each anchor frame is a training sample, and each anchor frame is either labeled as a background or associated with a real border. Each anchor box predicts an edge box. For each anchor box predict its type and predict the edge box. In YOLO algorithm, for different datasets, there are anchor boxes with initial set length and width. In network training, the network outputs a prediction frame based on the initial anchor frame, which in turn is compared to the real frame, calculates the gap between the two, and then updates it in reverse to iterate the network parameters. This functionality is embedded into the code in YOLOv5, which adaptively calculates the optimal anchor frame value for different training sets each time it is trained. The best anchor frame value is output using non-maximal suppression (NMS), which merges similar predictions, selects the largest predicted value that is a non-background class, and removes all other predictions that have an IoU value greater than x with it. NMS specifically means that first, the target bounding box lists are sorted according to their corresponding confidence scores; Next, select the highest confidence ratio bounding box to add to the final output list and remove it from the bounding box list; Next, calculate the area of all bounding boxes; then calculate the IoU of the highest confidence bounding box with respect to the other candidates and remove the bounding boxes with an IoU greater than the threshold value. (A threshold is set based on the list of target bounding boxes and their corresponding confidence score lists, and the threshold is used to remove bounding boxes with large overlaps) Finally, repeat the process until the bounding box list is empty.

2.2.4. Multiple Convolutional Layer Blocks to Compress Images. The image gets smaller and smaller as it is compressed by multiple blocks of convolutional layers, allowing the bottom segment to fit small objects and the top segment to fit large objects[6]. In commonly used target detection algorithms, different images have different lengths and widths, so a common approach is to uniformly scale the original image to a standard size before feeding it into the detection network. For example, the YOLO algorithm is commonly used in the 416*416, 608*608 and other sizes to scale the following 800*600 image. However, this has been improved in YOLOv5 code, and YOLOv5's inference can be very fast, which is a good improvement. We believe that in the actual use of projects, many images have different aspect ratios, so after zooming in and out for images and filling, the blach border on both sides of the image is a different size. And if too much is filled in, there is redundancy of information that slows down reasoning. As a result, the letterbox function in the YOLOv5 has been improved so that a minimal black border can be added to different original images depending on the situation. The black edges at the ends of the image height decrease, and during inference, the computational complexity also decreases, resulting in an improvement in target detection speed. Step 1: Calculate the original scaling size of the scaling ratio, separate the initial image sizes, and select a smaller scaling factor. Step 2: Calculate the scaled size, the smallest scaling factor of 0.52 are multiplied by the width and length of the image, and the width becomes 416, while the height becomes 312. Step 3: Calculate the fill value needed for the black border (416-312=104),so that the initial height of the image to be filled can be obtained. Then, by taking the remainder of np.mod in numpy, we obtain 8 pixels and divide them by 2 to obtain the values that need to be filled at both ends of the image height. In YOLOv5's algorithm, the above operating steps need to be repeated several times[7].

DOI: 10.54254/2755-2721/50/20241583

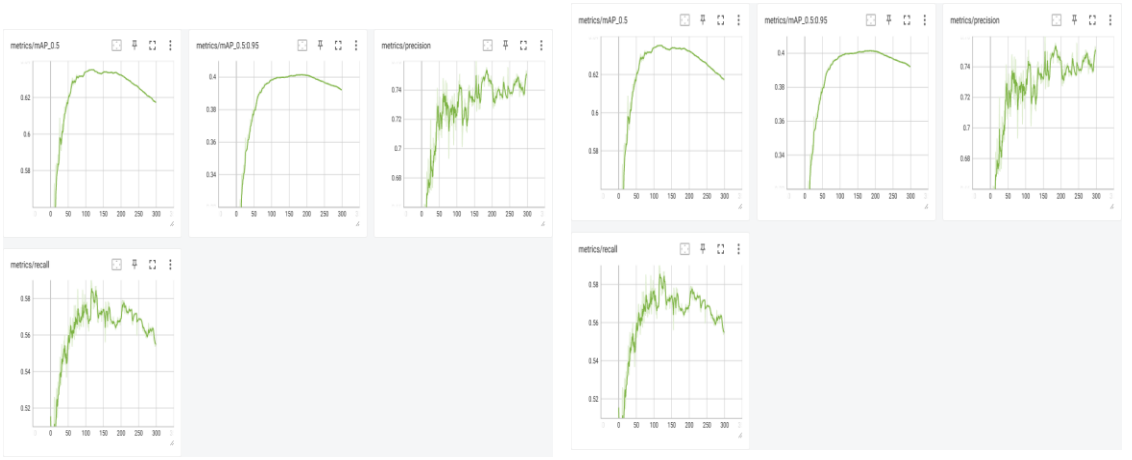
2.3. Model training

Before training, the adaptive anchor frames will automatically calculate the best recall of the default anchor frames based on the number of dataset annotation information provided, by calculating the best recall the model will be more likely to capture targets of different sizes and shapes, reducing the possibility of missed detection. At the same time, calculating the optimal recall can determine the number of anchor frames to be learned more precisely and avoid unnecessary anchor frames, thus reducing the complexity of the model and the risk of overfitting, and helping to improve the generalization ability of the model. When targeting real-time video analytics, reducing the number of anchor frames that need to be evaluated saves computational resources and time and improves system performance.In the training process of the images, 80% of the images are putted into training and 20% intovalidation.Data is preprocessed and enhanced through the use of official pre-trained models: First,adjust the image size to 640*640, and the processed image and the label corresponding to the image are obtained by mosaic data enhancement.After concatenating and enhancing the data, the image is input into the neural network. Figure 2 shows some example images after data enhancement [8].

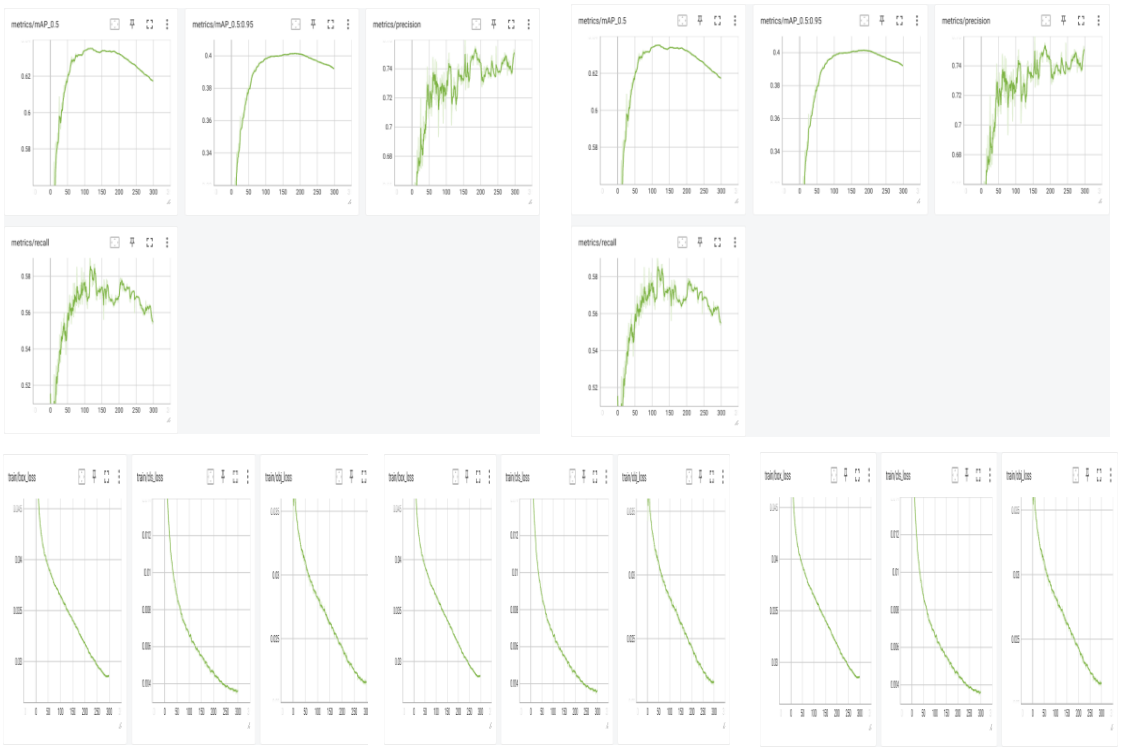
3. Experiment and performance

3.1. Quantitative comparison

We performed two separate training sessions using the same dataset, Figure 1 (a) shows the results of training with 300 epochs and Figure 1 (b) shows the results of training with 800 epochs. It can be noticed that in the phase when the model training is just started, it performs poorly in detecting some categories of targets because the parameters of the model are randomly initialized. This leads to a relatively low and very low mAP value in this phase. As the training progresses, the model gradually learns a better representation of the features is and the mAP value rises. Different categories and samples have different importance, certain categories appear relatively frequently in the dataset, which leads to these targets being more easily detected by the model, so the model is more likely to perform well on these categories at the beginning, which leads to an increase in the mAP value, and the mAP value decreases as the model focuses on more challenging categories as the training progresses. Also we notice that the values of mAP for the model that underwent 800 rounds of training decreased more significantly relative to the values of mAP for the model that underwent 300 rounds of training, when the number of training rounds is high, the model has more chances to learn the specific examples and noises in the training data, which results in a model that performs very well on the training data but not on the new data. When the number of training sessions increases, the model has more chances to memorize the training data, which means that the capacity of the model (i.e., the ability to learn and fit the data) also increases. If the capacity of the model is too large, it may overfit the training data instead of generalizing to unseen data.

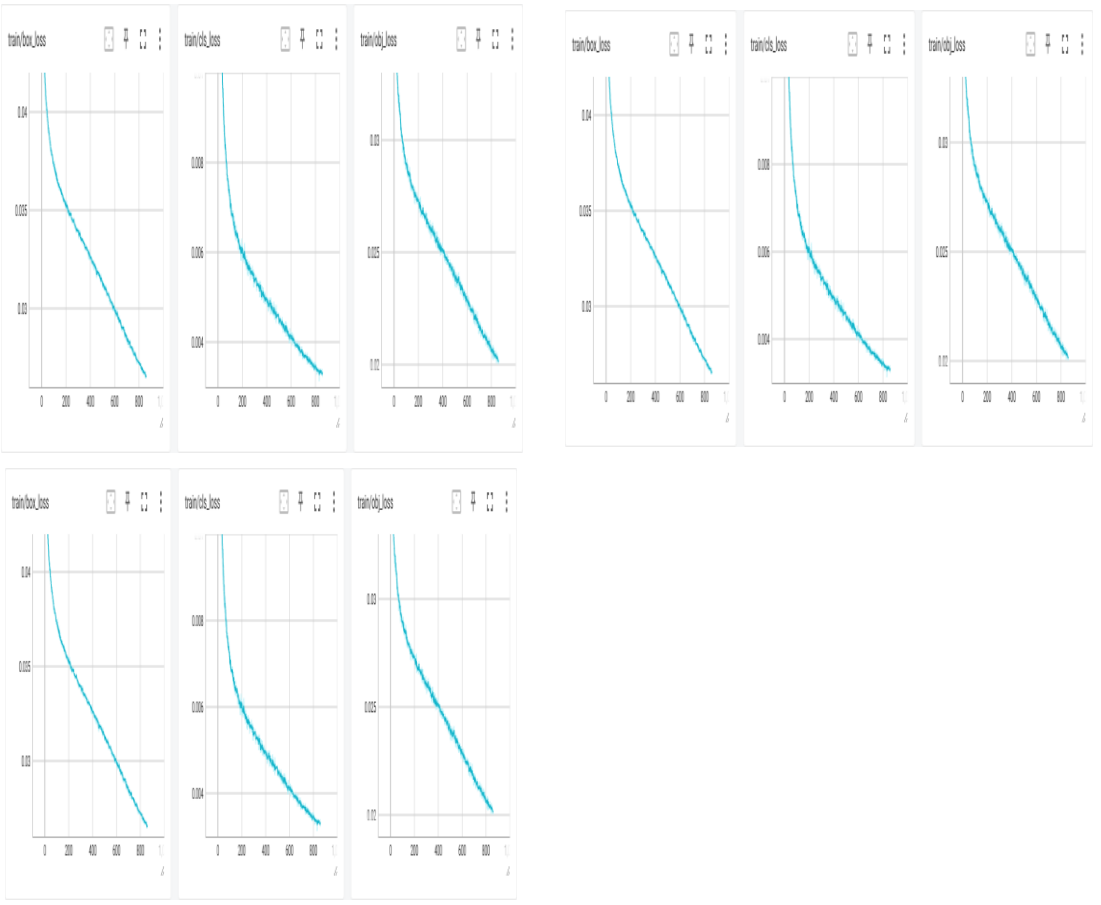


DOI: 10.54254/2755-2721/50/20241583



(a) Loss and accuracy of training with 300 epochs

DOI: 10.54254/2755-2721/50/20241583



(b) Loss and accuracy of training with 800 epochs

Figure 1. Performance comparison with models trained with different epochs

Accuracy is one of the metrics used to evaluate the performance of a classification model, and it indicates how many of the samples predicted by the model to be positive categories are actually positive categories. Precision rate rises to its highest value between the number of 300 training rounds and the number of 500 training rounds, but as the model becomes more confident, it leads to more false-positive errors, which leads to a decrease in precision rate. Recall reaches its highest value between 100 training rounds and 200 training rounds. The reason for this could be that the difference in the number of samples between the positive and negative categories in the dataset is too large, which leads to the fact that the model may be more likely to predict the negative categories early in the training.[9] As the model is trained, it may begin to notice positive categories and recall increases. Subsequently the model may i.e. miss some of the true positive categories, leading to a decrease in recall. The data shows that the Loss function continues to decrease, which usually indicates that the model is gradually improving its performance on the training data. At the same time the Loss function is gradually stabilizing and the model may be close to convergence.

3.2. Visualization analysis

We also visualize the detection results of different models in Figure 2. The experimental results show that a higher number of training rounds helps in the detection of categories with less sample data.

DOI: 10.54254/2755-2721/50/20241583

DOI: 10.54254/2755-2721/50/20241583

(a) Detection training with 300 epochs (b) Detection training with 800 epochs

Figure 2. Comparison of visualization results with models trained with different epochs

Figure 3 further shows the number of labels detected in the model. The data shows that the model with 800 rounds of training detects the 'car' label with higher confidence than the model with 300 rounds of training, which leads to higher precision and higher confidence of the model on the labels in complex environments[10] with a higher number of training rounds. However, more training rounds can lead to a decrease in the recall of the model, which will result in the model missing targets that need to be inspected during the detection process. Meanwhile, among all the labels 'car' has the most amount of labels and 'bus' has the least amount of labels, according to the results of the test, more training times can improve the model's confidence in the labels with the same data set.

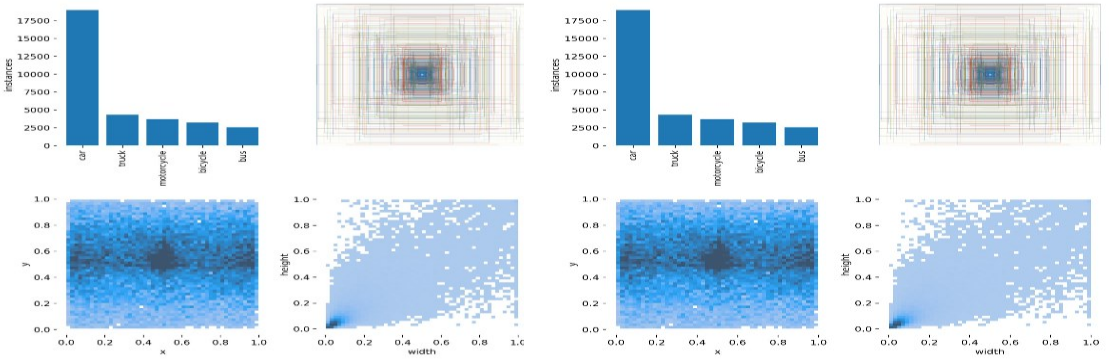


Figure 3. Statics of number of labels detected in the model

4. Discussion

In recent years, vehicle recognition has been widely used in daily life and has great significance. Vehicle recognition technology is an intelligent technology based on computer vision and image processing technology. It can achieve automatic recognition and classification of vehicles by analyzing and processing vehicle images. Vehicle recognition technology has extensive applications and important significance in fields such as traffic management, safety monitoring, and intelligent transportation.

Vehicle recognition technology has played an important role in traffic management. By identifying and classifying vehicles, functions such as traffic flow statistics, vehicle type analysis, and vehicle illegal behavior monitoring can be achieved, providing important data support and decision-making basis for traffic management departments. For example, in the case of urban traffic congestion, traffic management departments can use vehicle recognition technology to monitor the traffic flow on the road in real time, adjust the time of traffic lights in a timely manner, optimize traffic flow, and alleviate traffic congestion.

DOI: 10.54254/2755-2721/50/20241583

Vehicle identification technology also plays an important role in safety monitoring. By identifying and tracking vehicles, it is possible to monitor and record their driving trajectory, dwell time, parking location, and other information, providing important data support and basis for safety monitoring departments. For example, in urban security monitoring, vehicle recognition technology can achieve real-time monitoring and tracking of vehicles, timely detection and handling of safety incidents such as traffic accidents and illegal activities, ensuring the safety and stability of the city.

Nowadays, there are still a lot of problems in vehicle and pedestrian detection, such as: the speed of machine recognition needs to be improved, how to recognize vehicles and pedestrians when intentionally disguised, how to clearly recognize each object even when there is a large overlap of detection targets, and so on.

In the future, vehicle and pedestrian detection will enter more fields, such as our group's research on vehicle and pedestrian detection in unmanned areas, how to improve recognition and discrimination rates, and eliminate various external interference factors. In the future, more efficient vehicle and pedestrian recognition can improve people's quality of life and safety level.

5. Conclusion

Through the YOLOV5 algorithm to implement the object detection, its accuracy rises greatly, experiments show that after 300 rounds of training rectangular box loss value gradually slides to about 0.0325 tends to stabilize, confidence loss value gradually slides to about 0.0035 tends to stabilize, classification loss value gradually slides to about 0.0215 tends to stabilize. It is shown that YOLOv5 has good convergence to homemade datasets. And experimentally we found that mAP (IOU set to 0.5) had the highest value at two hundred training sessions, around 0.64, and mAP (IOU from 0.5 to 0.95) was above 0.4. In summary, with the advantages of adaptive anchor frame calculation and adaptive image scaling in yolov5, the accuracy of vehicle recognition is greatly improved. And, by increasing the number of training rounds, the confidence of car labeling increases, but it is worth noting that as the number of training rounds increases, the recall of the model is lower, which makes the model miss some targets that need to be detected. Vehicle-based object detection has a wide range of applications, such as object detection that can be applied in uninhabited areas. The unmanned rover can realize the environmental perception and object recognition of no man's land through advanced radar, infrared, lidar and other technical means, so as to detect illegal intruders in time. In addition, the unmanned rover can also achieve autonomous obstacle avoidance and safe driving through autonomous decision-making and path planning, improving detection efficiency and safety. Object detection can also be used for medical diagnosis. Identify cancerous tissue by object detection at a level comparable to that of a trained physician. In pathology, machine vision can augment efforts traditionally reserved for microscopic pathologists. And facial recognition software combined with object detection can help clinicians diagnose rare diseases. Patient photos are analyzed using facial analysis and deep learning to detect manifestations associated with rare genetic diseases. So, object detection has a wide range of applications and can greatly improve our quality of life.

Authors contribution

All authors contributed equally to this research, and their names are listed in alphabetical order.

References

[1] Luo Jianghong, Yuan Ziqi, Yi Zhixiong. Based on the improved YOLOv5 urban road vehicle pedestrian detection method [J]. Electronics, 2023,31 (19): 67-72.

[2] Zhang Shuqing, Wang Yachao, Xiao Han. Improved YOLOv5 based pedestrian and vehicle detection algorithm in haze weather [J/OL]. Radio Engineering 1-10 [2023-10-22]

[3] Jiang Chao, Zhang Hao, Zhang Enze et al. Pedestrian and vehicle object detection algorithm based on improved YOLOv5s [J]. Journal of Yangzhou University (Natural Science Edition), 2022-25 (06): 45-49.

DOI: 10.54254/2755-2721/50/20241583

[4] Shea K, Nash R. An introduction to convolutional neural networks[J]. arXiv preprint arXiv:1511.08458, 2015.
[5] Zhang A, Lipton Z C, Li M, et al. Dive into deep learning[J]. arXiv preprint arXiv:2106.11342, 2021.
[6] Yan Zhao, Xiangwei Kong, Chunbin Ma, etc. Real-time circuit board fault detection algorithm based on Darknet network and YOLO4 [J]. Computer Measurement and Control, 2023, 31(06):101-108.
[7] Bochkovskiy A, Wang C Y, Liao H Y M. Yolov4: Optimal speed and accuracy of object detection[J]. arXiv preprint arXiv:2004.10934, 2020.
[8] Fan Z, Zhu Y, He Y, et al. Deep learning on monocular object pose detection and tracking: A comprehensive overview[J]. ACM Computing Surveys, 2022, 55(4): 1-40.
[9] Yongjie Ma, Yunting Ma, Shisheng Cheng, et al. Road vehicle detection method based on improved YOLOv3 model and Deep-SORT algorithm [J]. Journal of Transportation Engineering, 2021, 21(2): 222-231.
[10] Yao Guan, Kai Zhu. Target detection on traffic roads under foggy conditions [J]. Computers and Telecommunications, 2023, 1(5): 69.