information

MDPI

Article

Transfer Learning-Based YOLOv3 Model for Road Dense Object Detection

Chunhua Zhu [1,2,3,*], Jiarui Liang [1,2,3] and Fei Zhou [1,2,3]

[1] Key Laboratory of Grain Information Processing and Control, Henan University of Technology, Ministry of Education, Zhengzhou 450001, China; 2021920294@stu.haut.edu.cn (J.L.); f.zhou@haut.edu.cn (F.Z.)

[2] Henan Key Laboratory of Grain Photoelectric Detection and Control, Henan University of Technology, Zhengzhou 450001, China

[3] College of Information Science and Engineering, Henan University of Technology, Zhengzhou 450001, China

* Correspondence: zhuchunhua@haut.edu.cn

Abstract: Stemming from the overlap of objects and undertraining due to few samples, road dense object detection is confronted with poor object identification performance and the inability to recognize edge objects. Based on this, one transfer learning-based YOLOv3 approach for identifying dense objects on the road has been proposed. Firstly, the Darknet-53 network structure is adopted to obtain a pre-trained YOLOv3 model. Then, the transfer training is introduced as the output layer for the special dataset of 2000 images containing vehicles. In the proposed model, one random function is adapted to initialize and optimize the weights of the transfer training model, which is separately designed from the pre-trained YOLOv3. The object detection classifier replaces the fully connected layer, which further improves the detection effect. The reduced size of the network model can further reduce the training and detection time. As a result, it can be better applied to actual scenarios. The experimental results demonstrate that the object detection accuracy of the presented approach is 87.75% for the Pascal VOC 2007 dataset, which is superior to the traditional YOLOv3 and the YOLOv5 by 4% and 0.59%, respectively. Additionally, the test was carried out using UA-DETRAC, a public road vehicle detection dataset. The object detection accuracy of the presented approach reaches 79.23% in detecting images, which is 4.13% better than the traditional YOLOv3, and compared with the existing relatively new object detection algorithm YOLOv5, the detection accuracy is 1.36% better. Moreover, the detection speed of the proposed YOLOv3 method reaches 31.2 Fps/s in detecting images, which is 7.6 Fps/s faster than the traditional YOLOv3, and compared with the existing new object detection algorithm YOLOv7, the speed is 1.5 Fps/s faster. The proposed YOLOv3 performs 67.36 Bn of floating point operations per second in detecting video, which is obviously less than the traditional YOLOv3 and the newer object detection algorithm YOLOv5.

Keywords: dense road; object detection; darknet-53 network; transfer learning

check for updates

Citation: Zhu, C.; Liang, J.; Zhou, F. Transfer Learning-Based YOLOv3 Model for Road Dense Object Detection. *Information* **2023**, *14*, 560. https://doi.org/10.3390/info14100560

Academic Editor: Marco Leo

Received: 24 August 2023
Revised: 5 October 2023
Accepted: 10 October 2023
Published: 12 October 2023

1. Introduction

With the continuous development of modern society, road and bridge construction has a considerable scale, and the number of vehicles in the city is also increasing, which has led to frequent traffic accidents in recent years; therefore, the need for better management and planning of vehicles on the road, road vehicle detection technology, came into being. Through this road vehicle detection technology, traffic regulation personnel can more efficiently monitor the violation behavior and can effectively administrate the road pipeline. At the same time, road vehicle detection technology can be used in the construction of intelligent transportation, through that technology, can achieve automatic driving, automatic parking, automatic identification and other functions, improve the efficiency of urban traffic operations, and also bring convenience to people's driving.

The dense detection of objects, such as vehicles and pedestrians, on road scenes faces several problems: object overlapping or occlusion, uneven distribution of objects, and difficulty in detecting edge objects. Traditional object detection techniques may be separated into the following two broad categories: R-CNN (Region-Convolutional Neural Network), which is based on candidate regions, including R-CNN [1], Fast R-CNN [2], Faster R-CNN [3], and other two-stage networks; and regression-based single-stage networks, like YOLO [4], SSD (Single Shot MultiBox Detector) [5]. R-CNN algorithms need to extract the characteristics of the candidate regions before feeding them into a pre-trained CNN model in order to obtain a characteristic for classification. Due to the large amount of object overlapping in the candidate regions, the features from the overlapping regions will be repeatedly calculated when extracting features, thereby the real-time performance is poor; comparably, the YOLO model only requires one feed-forward neural network to directly predict the classifications and positions of diverse objects for several independent candidate regions, which has the advantages of simpler training, better detection efficiency, etc [6]. YOLO has been widely used in object detection [7], and it has evolved to become the YOLOv3. The subsequent YOLO series such as YOLOv5 and YOLOv7 are all improved on the basis of YOLOv3 [8]. Compared to the common target-detection algorithms, the YOLOv3 model has the stronger self-learning ability and larger capacity, which can solve the problem of poor accuracy from the excessive high-resolution dataset and unbalanced positive and negative samples; additionally, the YOLOv3 has more candidate frames and can improve the intersection ratio during detection. For matching the special road dense object detection scene, transfer learning [9] will be introduced to the YOLOv3 network to fine-tune and accelerate the original training model, which is embedded in the output layer of pre-trained YOLOv3 and, thereby, one new transfer learning-based YOLOv3 model is shaped.

The key contributions can be described as follows.

(1) Based on the pre-trained YOLOv3 model, the transfer training is introduced as the output layer for the special dataset containing detecting objects; moreover, one random function is adapted to initialize and optimize the weights of the transfer training model, which is separately designed from the pre-trained YOLOv3.

The transfer learning used in this paper first calculates the eigenvector of the convolution layer of the pre-trained model to all the training and test data, then freezes the convolution layer of the pre-trained model and trains the remaining convolution layer and the fully connected layer. There are three advantages to this operation. First, the initial performance of the model is higher before the model parameters are fine-tuned. Secondly, in the training process, the model improves faster. Finally, after the training, the obtained model converges better.

(2) In the proposed YOLOv3, the object detection classifier replaces the full convolutional layer of the traditional YOLOv3, aiming to relieve the conflict distinguishing the edge target features and other background features caused by the excessive receptive field of the full connection layer; additionally, the introduced classifier can avoid the excessive computation from the fully connected layer in the traditional YOLOv3.

In the traditional YOLOv3 network, the convolutional layer before the fully connected layer is responsible for image feature extraction. After obtaining features, the traditional method attaches the fully connected layer before activating classification. The idea of the object detection classifier is to use average global pooling to replace the fully connected layer, and more importantly, the empty space extracted by the previous convolutional layer and, thus, the pooling layer is preserved. Intermediate and semantic information is used so that the detection accuracy will be improved in practical applications. In addition, the object detection classifier removes the limit on the input size and has essential applications in the process of convolutional visualization.

2. Related Works

Road vehicle detection adapts the advanced technology and equipment in order to realize the comprehensive monitoring and detection of vehicles running on the road.

Through various sensors, cameras, and computer systems, the relevant information of vehicles can be collected and processed in real time, providing essential data support to road management departments and transportation enterprises and effectively improving road safety and traffic management levels. By surveying the existing road vehicle detections, we can see that the traditional methods mainly include infrared detection, ultrasonic detection, radar detection, and object detection.

2.1. Infrared Detection

The infrared detection method consists of a modulated pulse generator that generates a modulation pulse through the infrared probe radiation to the road; when there is a vehicle, the infrared pulse from the vehicle's body is reflected by the probe-receiving tube. This method is more dependent on the optical environment, and dust particles in the light will affect the routine work of the system. Ai Hong proposed the overall scheme, including the hardware and software design of the vehicle flow monitoring system based on an infrared sensor [10]. Furthermore, the section of the infrared sensor and the application of the Fresnel lens are explained in detail. After two amplification stages, the infrared sensor is compared with the window comparator to generate pulses into the MCU circuit; thereby, the number of passing vehicles is obtained by analyzing and judging the pulse sequence. The experimental results show that the proposed detection system runs fast, has a low false-alarm rate, and has strong reliability.

2.2. Ultrasonic Detection

Ultrasonic detection considers the influence of a vehicle's shape on ultrasonic. However, because ultrasonic has a certain diffusion angle, a single ultrasonic can only measure the distance and cannot measure the orientation, so multiple ultrasonic sensors from multiple angles can be chosen; however, they have a larger cost and are unsuitable for daily scenes. Zhao Yani proposed the ultrasonic vehicle detector [11], which consists of three parts: an ultrasonic probe, a host, and communication; here, the data from the probe is directly analyzed, processed, and stored and then sent back to the central processor. The experimental results show that the detection accuracy of this detector can reach 99%, the average speed detection accuracy can reach 90%, and the accuracy of the model classification can reach 94%.

2.3. Radar Detection

Radar detection recognizes vehicles by emitting microwaves toward the road surface and receiving the reflected waves, but it can only detect vehicles that are in motion and must also use a unique sensor design or signal processing software to detect stationary vehicles. Jin Lu proposed a method of vehicle detection based on millimeter-wave (MMwave) radar and visual multi-features [12], which is implemented by three steps. Firstly, a spatial alignment algorithm is proposed to realize the spatial alignment of MMwave radar and vision; then, the region of interest of the target vehicle is extracted according to the results of spatial alignment and searching strategy; finally, vehicle detection is realized by integrating features such as shadows under the car, symmetry axis, and left and right edges. The method is proved the efficiency and reliability of the proposed radar vehicle detection.

2.4. Object Detection

With the rise of autonomous driving technology, deep learning-based object detection technology has also developed, which usually comprises an electronic camera, image processor, and display parts. The camera takes a continuous recording in a certain area of the road, and in the image processor, the images are transformed into a digital signal. Then, the microprocessor processes the image background to identify the existence of the vehicle and the vehicle type. This method can detect multiple traffic parameters and has a wide detection range with better real-time capability and flexibility, which has obtained a wide application in road vehicle detection. Wang Tingting proposed the recursive YOLOv4 target detection based on an attention mechanism, namely the RC-YOLOv4 algorithm [13].

RC-YOLOv4 combines the channel and spatial attention mechanism, aiming to help the network model pay more attention to the key information and small target information in the detected image. The experimental results show that the average precision of the proposed RC-YOLOv4 algorithm is 12. 69% higher than the YOLOv4 in the self-made vehicle detection data set. Xiong Liyan proposed vehicle detection based on lightweight MobileVit [14]. This method adapted the MobileVit network as the backbone feature extraction network of the model in order to fully extract the feature information and make the model lightweight; in the prediction layer network, multiscale vehicle detection and recognition based on the PANet network is presented to improve the detection performance of the model for small targets. The experimental results show that the average detection accuracy of this algorithm is 98.24%, the detection speed is 58 ms per picture, and the comprehensive performance is better compared with the other algorithms. Yuan Lei proposed the improved road object detection as CTC-YOLO (context transformer and convolutional block attention module based on YOLOv5) [15], which can increase the detection accuracy of small targets by improving the network detection head and adding the multi-scale target detection layer. The experimental results show that the proposed CTC-YOLO reaches 89.6%, 46.1%, and 57.0%, on the publicly available datasets KITTI, Cityscapes, and BDD100K, respectively. A. Alshehri et al. proposed the Residual Neural Networks for Origin–Destination Trip Matrix Estimation from Traffic Sensor Information [16]; it deduces the correlation between traffic statistics and network topology from traffic characteristics. In order to train the proposed deep learning architecture, random synthetic flow data is generated from the historical demand data of the network. Large-scale networks are used to test and validate the performance of the models. Xue Q W et al. proposed a three-dimensional vehicle detection algorithm based on multi-modal feature fusion [17]; by fusing the point cloud and image information with the average pixel value, the feature pyramid is added to extract the fused advanced feature information, improve the detection accuracy in complex road scenes, establish the feature fusion region recommendation structure, and generate the region recommendation based on the advanced feature information. Wang W H proposed that the subway obstacle detection system is designed based on infrared and visible light imaging systems [18]. Based on the functional requirements analysis and operating platform, the system function modules and hardware components are defined. Due to the different imaging mechanisms of infrared and visible image sensors, there are also some differences in the captured images. The Bi-dimensional integrated Empirical Mode decomposition (BEEMD) algorithm is used to fuse the image, and the YOLO network is used to detect the category of obstacles.

Table 1 summarizes these recently-emerged target detection algorithms for vehicle detection and the Data set, Key aspects, Advantage, mAP, FPS, and Platform of the proposed YOLOv3 are also provided.

As can be seen from Table 1, the proposed YOLOv3 can improve the detection speed without the loss of detection accuracy, compared with the recently-emerged existing vehicle detection algorithms.

Table 1. Comparison among the different algorithms.

Algorithms	Data Set	Key Aspects	Advantage	mAP	FPS	Platform
RC-YOLOv4	self-made vehicle detection data set	Combine the channel and spatial attention mechanism	Pay more attention to the key information and small target information	87.33%	43	Ubuntu 18.04
MobileVit	KITTI	Introduce MobileVit network as the backbone feature extraction network	Fully extract the feature information and make the model lightweight	95.36%	58	Window 10

Table 1. *Cont.*

Algorithms	Data Set	Key Aspects	Advantage	mAP	FPS	Platform
CTC-YOLO	KITTI	Add the context transformer and convolutional block attention module	Increase the detection accuracy of small targets by improving the network detection head and adding	89.6%	69	Window 10
The proposed YOLOv3	KITTI	Adapt the transfer learning and replace the detection head of the original network	While ensuring the detection accuracy, the training and detection time is reduced	88.24%	78	Window 10

2.5. Characteristics of Vehicle Detection Sensors

The characteristics of various vehicle detection sensors are summarized in Table 2.

Table 2. Characteristics comparison of traffic sensors.

Methods	Advantage	Malpractice
Infrared detection	Solve the day–night conversion problem, and provide a lot of traffic management information	An excellent infrared focal plane detector may be required, which requires increased power and reduced reliability to achieve high sensitivity
Ultrasonic detection	Small size, easy to install; Long service life; Movable; Multi-lane detection is possible	The performance decreases with the influence of ambient humidity and air flow
Radar detection	Work in all weather, keep excellent performance harsh weather; Detect static vehicles; Detect multiple lanes in lateral mode	The higher detector installation requirements because detection accuracy decline when the road has an iron strip
Object detection	Provide visual images for incident management; Provide a lot of traffic management information; A single camera and processer can detect multiple lanes	Smaller vehicles cannot be detected when obscured by large vehicles

3. Algorithm Principle

Figure 1 shows the architecture of the dense road detection system based on YOLOv3, which includes three parts: the YOLOv3 backbone network, a transfer training unit, and optimization of network parameters. Firstly, the VOC 2007, VOC 2012, and COCO datasets are selected for YOLOv3 network pre-training; then, the images containing vehicles are extracted from the VOC 2007 dataset and re-labeled to form the special vehicle dataset, and the transfer training-based YOLOv3 model is transferred and trained in this dataset; finally, the test dense road pictures or videos are input into the proposed model, and the output is obtained by feature extraction, multi-scale detection, and non-maximum suppression (NMS) processing. Performance evaluations are performed by confidence, mean precision (mAP, mean Average Precision), and precision recall (P-R, Precision Recall) of object detection.

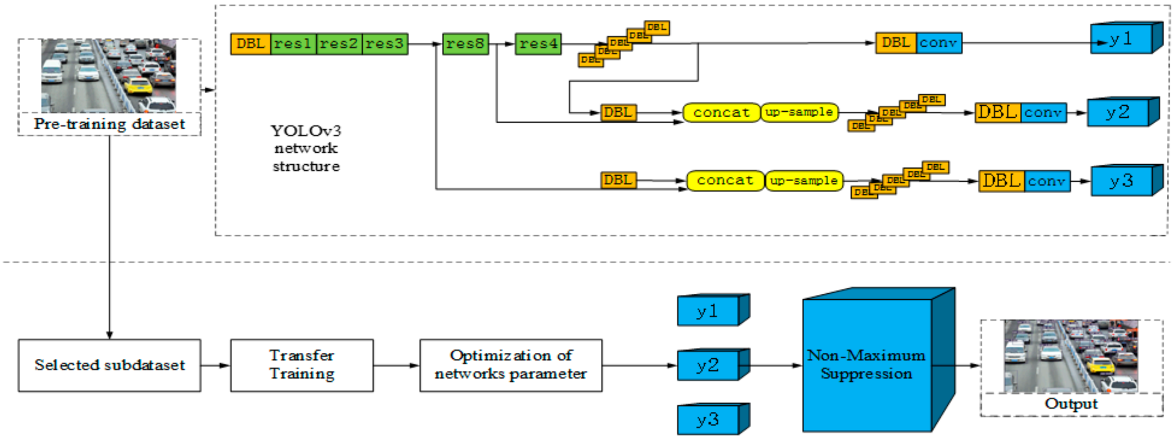


Figure 1. Architecture of dense road object detection model.

3.1. Feature Extraction Network

Compared with the YOLOv2 network, the backbone portion of the YOLOv3 network has evolved from Darknet-19 to Darknet-53, consequently expanding the number of network layers [19] and adding the cross-layer sum operation in the residual network; YOLOv3 network has fifty-three convolutional layers (ResNet, Residual Network). Darknet-53 is an entirely convolutional network comprised of 3 × 3 and 1 × 1 convolutional layers, including 23 residual modules and layers of detection channels that are completely interconnected. As depicted in Figure 2, the convolutional layers are interconnected by quick link [20] (that is, SC, Shortcut Connections). This SC structure can greatly enhance the computation performance of the network, enabling the network to obtain faster detection speed in a limited number of network layers. In the detection architecture, YOLOv3 separates three channels for feature detection into distinct grid sizes. These channels include feature maps with grid sizes of 52 × 52, 26 × 26, and 13 × 13, which correspond to the detection of large-scale (y1), medium-scale (y2), and small-scale (y3) picture features, respectively. Thereby, The YOLOv3 can provide a higher detection accuracy with fewer network parameters and fewer superfluous network layers, enabling it to improve both the detection speed and the detection accuracy. By comparison, the conventional R-CNN relies on deepening the network structure to enhance the recognition rate.

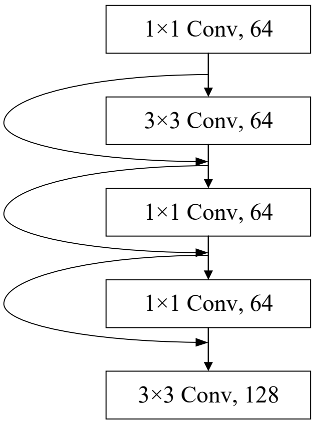


Figure 2. Schematic diagram of quick link.

3.2. Training Strategy Design

The traditional deep learning network generally improves the recognition accuracy by increasing the training set or deepening the network complexity. In the application of road dense object detection, the real-time detection is an important indicator [21]. Transfer learning is introduced in an effort to improve the training process of the traditional YOLOv3 network, as shown in Figure 3.

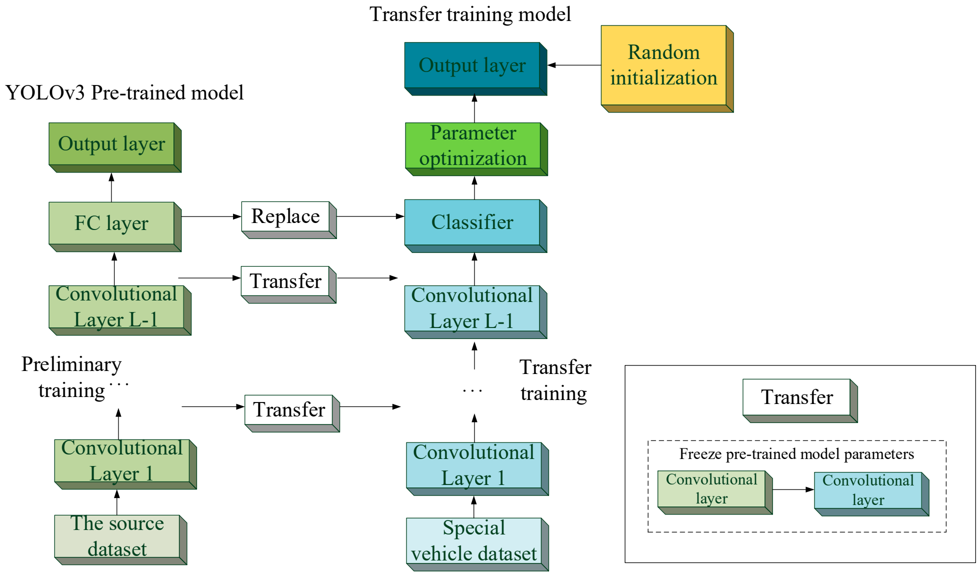


Figure 3. Transfer training strategy.

In Figure 3, the YOLOv3 backbone network is combined with the transfer training model; during the pre-training process, the YOLOv3 network is trained in the VOC 2007 and VOC 2012 datasets to obtain the pre-trained model for 20 object types in the VOC dataset. Additionally, the COCO dataset is trained to obtain 60 object types, and a total of 80 object types can be obtained; then, the transfer training unit substitutes the convolutional layer of the YOLOv3 backbone network, its parameters are not fixed, which means they can be randomly initialized and optimized [22], on the special vehicle dataset. During pre-training, the epoch is set to 300, and the average precision (AP, Average Precision) and mAP of the transfer training model are shown in Table 3.

Table 3. Performance of YOLOv3 model.

Classes	AP	TP	FP	mAP
car	81.25%	79.65%	20.34%	
person	76.43%	60%	40%	
bicycle	74.12%	82.81%	17.18%	
motorcycle	73.08%	74.14%	25.85%	83.85%
bus	79.14%	81.06%	19.02%	
dog	72.44%	77.07	22.92%	
cat	73.34%	65.41%	34.58%	

In the transfer training model, the proposed YOLOv3 retains the convolutional layer in the traditional YOLOv3 after pre-training. When performing feature extraction, the pre-trained convolutional layers [23] are selected to be a feature extractor, the structure of which allows the input image to propagate forward. The convolutional layer parameters [24] retained from the pre-trained model are frozen and they take the output of convolutional layer L-1 as the proposed YOLOv3 extracted feature [25]. Such a convolutional layer can enable the network to better obtain the semantic information of the target and train it, thereby obtaining higher detection accuracy; then, the proposed YOLOv3 replaces the full convolutional layer of the traditional YOLOv3 with the object detection classifier [26], which can relieve the conflict of distinguishing between the edge target features and other background features caused by the excessive receptive field of the full connection layer; furthermore, it can reduce the problem of excessive computation caused by the fully connected layer, so that the proposed YOLOv3 can train and detect objects faster, and the corresponding network parameters are accordingly optimized [27] to increase dense detection precision on the road while simultaneously reducing training complexity.

In Table 2, AP represents the mean precision of each object type in the test set. The term "correctly classifying a positive example as a positive example" (abbreviated "TP") refers to the degree of precision achieved when accurately identifying a positive example, while "falsely classifying a negative example as a positive example" refers to incorrectly identifying a negative example as a positive example (abbreviated as "FP").

During the transfer training, nearly 2000 images containing vehicles are screened and labeled. This is a special dataset built [28] for dense object detection in road situations. Included among the road objects in the dataset are cars, people, bicycles, motorbikes, trucks, cats, and dogs. There are 7 types that appear frequently in the road scene, as shown Table 4, and the remaining 13 types are not common.

Table 4. Selected special dataset.

Class Name	Car	Person	Bicycle	Motorcycle	Bus	Dog	Cat
Total number	1434	1360	860	430	300	220	180

From Table 3, a total of 1434 labeled car images in the dataset are selected as a special dense road sub-dataset. The images are then divided into a training set and a test set in a 4:1 ratio [29]. In addition, the classes of the special dense road sub-dataset must be modified to 7 and the file path of "train", "valid", "names", and "backup" must be modified correspondingly for the transfer training model to correspond with the extracted special dense road sub-dataset. This is necessary for the transfer training model to correspond with the special dense road sub-dataset. Meanwhile, parameters Batch, LEARN-RATE and IOU in config.py need to be optimized with the following considerations:

(1) The Batch function is set to 8. It can make the network complete an epoch in a few iterations and reach a local optimal state while finding the best gradient descent direction [30]. Increasing this value will prolong the training time but will better find the gradient descent direction; decreasing this value may cause the training to fall into a local optimum, or not converge.

(2) The LEARN-RATE function is set from 10^{-4} to 10^{-6}. During the training procedure [31], the total number of training cycles generally determines a learning rate that continually adapts to new input. Ten rounds of training are set. Experiments indicate that the learning rate is initially set at 0.0001 at the beginning of training and then gradually slows down after a certain number of rounds have been completed. As the training nears its conclusion, the learning rate is lowered until it hits 0.000001. The setting of the learning rate, which is based on ten training rounds, solves the problems of easy loss value explosion and easy oscillation caused by a learning rate that is too large at the beginning of training; if it is too small, it is easy to over-fit, resulting in slow network convergence.

(3) The IOU function is allocated the value 0.65. In computer detection tasks, the IOU value is equal to one and the intersection is the same as the union if the actual bounding box and the predicted bounding box entirely overlap [32]. Generally, a value of 0.5 is utilized as the threshold to determine whether or not the predicted bounding box is correct. It is possible to increase the detection accuracy of tiny items and edge objects while dealing with the detection of dense road objects by setting IOU to 0.65. This will enable the gathering of higher-quality samples and enhance the identification of dense road objects.

Next, we performed ablation experiments on three parameters in the model, as shown in Table 5.

Table 5. The results of ablation experiment.

Batch	Learn-Rate	IOU	Training Time/h	mAP (%)
			3.52	83.85
√			3.12	82.73
	√		3.37	83.14
√	√		2.93	82.93
√	√	√	2.96	87.64

Based on Table 4, setting the batch function to 8 can reduce the training time of the network by 0.4 h; setting the LEARN-RATE function from 1e-4 to 1e-6 can reduce the training time of the network by 0.15 h; by setting these two parameters to fixed values, the training time of the network can be reduced by 0.59 h, but the mAP is also reduced. Finally, the IOU value was set to 0.65, which reduced the training time of the network by 0.56 h, and the mAP value increased by 4.29%. Therefore, the three parameters in the network model are fixed and set, which reduces the training time of the network and improves the operation efficiency of the model under the condition of ensuring accuracy.

3.3. Loss Function Selection

In the proposed YOLOv3, the error loss mainly comes from the misjudgment of prediction frame, confidence, and category. The error in the prediction frame is determined by the rate of coincidence between the a priori frame and the prediction frame. If the rate at which the prediction frame and the a priori frame agree is large, this implies that the prediction is accurate and error margins are small; confidence error refers to the mistake induced by random sampling during testing in the test set, sampling the test set, and then estimating all test sets; category misjudgment error is the error caused by detecting one type into another. These three types of losses are expressed as *lbox*, *lobj*, and *lcls*, respectively.

$$lbox = \lambda_{coord} \sum_{i=0}^{s^2} \sum_{j=0}^{B} I_{i,j}^{obj} (2 - w_i \times h_i) [(x_i - \hat{x}_i) + (y_i - \hat{y}_i) + (w_i - \hat{w}_i) + (h_i - \hat{h}_i)] \tag{1}$$

$$lcls = \lambda_{class} \sum_{i=0}^{s^2} \sum_{j=0}^{B} I_{i,j}^{obj} \sum_{c \in classes} p_i(c) \log(\hat{p}_i(c)) \tag{2}$$

$$lobj = \lambda_{noobj} \sum_{i=0}^{s^2} \sum_{j=0}^{B} I_{i,j}^{noobj} (c_i - \hat{c}_i)^2 + \lambda_{obj} \sum_{i=0}^{s^2} \sum_{j=0}^{B} 1_{i,j}^{noobj} (c_i - \hat{c}_i)^2 \tag{3}$$

$$loss = lbox + lobj + lclcs \tag{4}$$

Here, s is the grid size; B is amount of prediction frames;$I_{i,j}^{obj}$ is the indicator function, which stand for if the prediction frame at i, j has an object, its value is 1, otherwise it is 0; $I_{i,j}^{obj}$ is the indicator function, which stands for if the prediction box is at i, j has no object and its value is 1, otherwise it is 0.

In this paper, ablation experiments were performed for each loss term in the loss function, and the experimental results are shown in Table 6.

Table 6. The results of the ablation experiment.

lbox	lobj	lcls	F1	mAP (%)
			57.31%	65.37
√			63.27%	73.64
	√		62.45%	72.82
		√	64.87%	75.68
√	√		69.39%	78.94
	√	√	68.76%	77.68
√		√	70.32%	79.56
√	√	√	71.47%	83.85

Based on Table 5, when the error in the prediction frame, confidence error, and category error are used separately, the F1 coefficient and mAP of the network model are low and cannot meet the requirements necessary for detecting road vehicles. When matching the three kinds of errors in pairs, it can be seen that the accuracy of the network model has been significantly improved; after the superposition of the three kinds of errors, the F1 coefficient and mAP in the network model reached 71.47% and 83.85%, respectively, which were able to meet the requirements of road vehicle detection in most scenarios. It can be seen from these seven groups of experiments that the error in the prediction frame and the category error greatly improve the detection effect of the network model, while the improvement effect of the confidence error is small.

4. Experimental Testing and Evaluation

4.1. Required Environment

The experimental configuration is shown in Table 7, using a Tensorflow (Developed by the Google team, America)framework in deep learning and an Opencv Python framework in computer vision.

Table 7. Configuration environment.

Operating System	CPU	Memory	GPU	CUDA	CUDNN
Windows 10(22H2)	Intel i5	8GB	NVIDIA GEFORCE RTX 745	CUDA 10.04	CUDNN 7.04

4.2. Detection Performance Evaluation

There are many evaluation criteria for target detection effect, among which the more common ones are intersection and combination ratio, P-R value, mAP, recall rate [33], accuracy, average accuracy rate, FPS (Frame Per Second), etc. IOU is the ratio of intersection and union of two rectangular frames, which is used to indicate the overlapping degree of rectangular frames A and B, the definition formula is as shown in Formula (5), and the accuracy, recall rate, and missing detection rate formula is as shown in Formula (7) to Formula (9).

$$IOU = \frac{|A \cup B|}{|A \cup B|} \tag{5}$$

$$precision = \frac{TP}{TP + FP} \tag{6}$$

$$recall = \frac{TP}{TP + FN} \tag{7}$$

$$AP = \sum_{i=1}^{n-1} (r_{i+1} - r_i) P_{inter}(r_i + 1) \tag{8}$$

$$mAP = \frac{\sum_{i=1}^{k} AP_i}{k} \tag{9}$$

Here, "true positive" (*TP*) refers to the genuine example, "false positive" (*FP*) refers to a false positive example, and "false negative" (*FN*) refers to a false negative example. Furthermore, the index of *mAP* [34] is also adapted to show the mean average precision. Where r_1, $r_2 ... r_i$ is the corresponding recall value at the first interpolation of the precision interpolated segment in ascending order. The average of the *AP* across all categories is the *mAP*.

Figure 4 displays the P-R curve generated by applying the proposed YOLOv3 to the data for the car object detection in the road scene. If the classification results of all test samples are positive, the recall rate of the model will be 1 and the accuracy rate will be extremely low; if the classification results of almost all test samples are negative, the accuracy rate will be high and the recall rate will be very low. Precision and recall are two rather contradicting metrics. Figure 4 shows that the proposed YOLOv3 is capable of reaching a high accuracy rate, that the recall rate slope drops in a moderate and steady fashion, and that the accuracy rate and recall rate achieve a more balanced state.

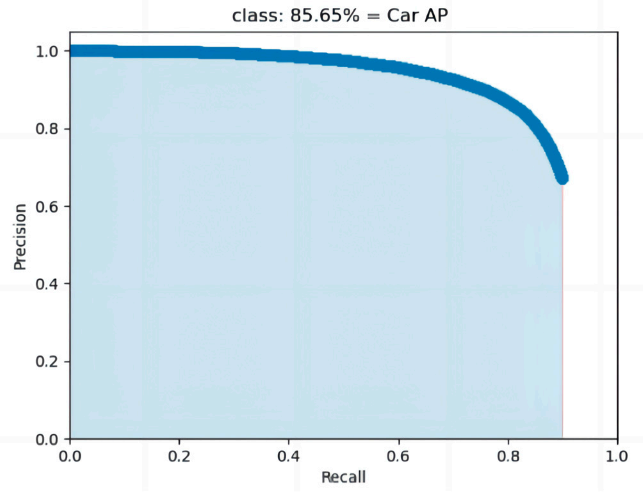


Figure 4. P-R curve of vehicle object detection.

In the special dense road sub-dataset, the training set and test set are divided into two sections with a 4:1 ratio. AP statistics are performed on the 7 types, as shown in Table 2, and *mAP* is calculated. Table 8 shows the performance comparison between several algorithms.

Table 8. Algorithm performance comparison.

Index / Algorithm	Param/M	FPS/Frames per Second	FLOPS/Floating-Point Operations per Second	Weight Size/M	mAP
YOLOv3	61.53	54.6	132.69 Bn	120.5	83.85%
YOLOv5	52.5	55.3	116.54 Bn	100.6	87.26%
YOLOv7	36.49	87.1	102.37 Bn	74.8	89.32%
The proposed YOLOv3	29.84	91.2	97.62 Bn	68.2	87.85%

In Table 7, The proposed YOLOv3 network model can achieve a relatively good performance, which is obviously superior to the 4% and 0.59% of the traditional YOLOv3 and YOLOv5, respectively; the number of parameters is 29.84 M, which is 6.65 M less than the latest YOLOV7 algorithm; the number of frames per second can reach 91.2 frames/s, which is obviously superior to the 4.1 frames/s of the latest YOLOV7 algorithm; the floating-point operations per second can reach 97.62 Bn, which is 4.75 Bn less than the latest YOLOV7 algorithm; the weight size is 63.2 M, which is 6.6 M less than the latest YOLOV7 algorithm. In conclusion, the proposed YOLOv3 algorithm reduces the size of the model, the amount of computation, and the number of parameters and also improves the detection speed of the network on the premise of ensuring the accuracy.

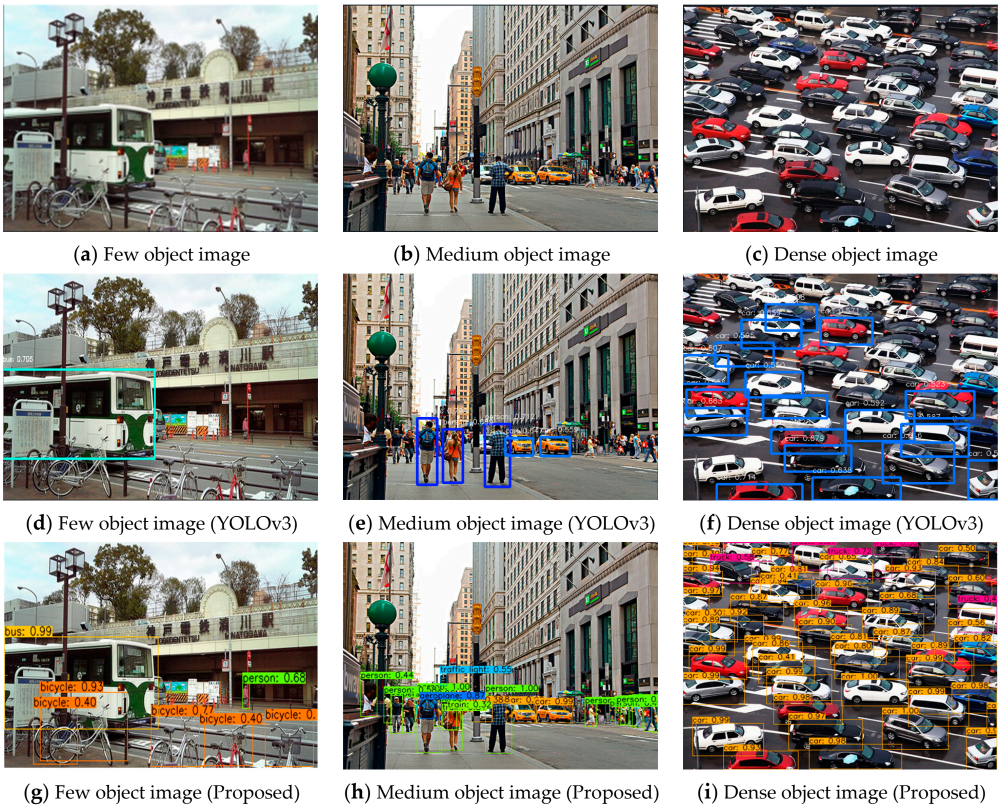
4.3. Comparative Analysis of Different Algorithms

Using the UA-DETRAC dataset as the test set, for detecting static road pictures, Table 9 shows the detection accuracy of each algorithm. Compared to the traditional YOLOv3 and YOLOv5, the proposed YOLOv3 approach exhibits an obvious improvement in the detection accuracy of each category and in the average detection accuracy.

Table 9. Comparison of the different algorithm performances.

Algorithm **mAP** **mAP**	**Car**	**Bus**	**Van**	**Others**	**Average mAP**
YOLOv3	79.61%	80.29%	76.21%	70.32%	75.1%
YOLOv5	83.32%	82.26%	79.64%	72.11%	77.87%
The proposed YOLOv3	82.73%	81.37%	80.36%	74.53%	79.23%

Figure 5 shows the original object image in the top row, the detection result of the traditional YOLOv3 in the second row, and the proposed YOLOv3 detection result in the third row. It is evident that the proposed YOLOv3 is capable of accurately differentiating between different types of vehicles and pedestrians, as well as accurately detecting pedestrians at the edge of the image; it is also capable of using a confidence box to indicate the object type division probability, and to mark the object type division probability. Additionally, the confidence will not decrease as the density of road objects increases.



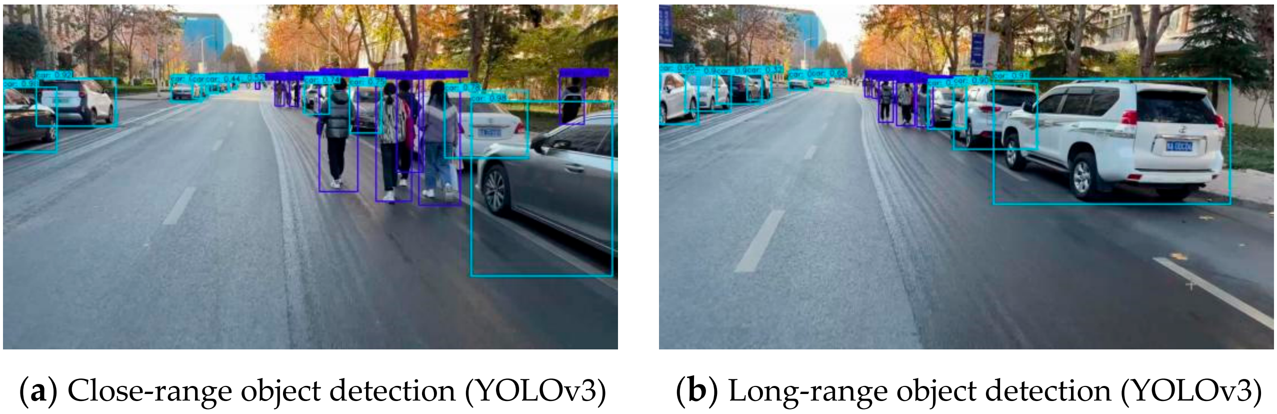
(**a**) Few object image (**b**) Medium object image (**c**) Dense object image

(**d**) Few object image (YOLOv3) (**e**) Medium object image (YOLOv3) (**f**) Dense object image (YOLOv3)

(**g**) Few object image (Proposed) (**h**) Medium object image (Proposed) (**i**) Dense object image (Proposed)

Figure 5. Object detection results of a single picture.

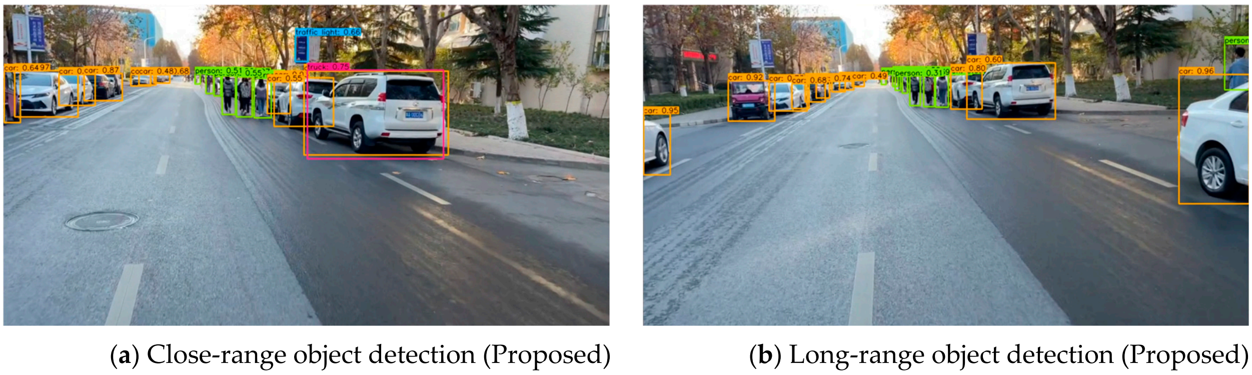
When detecting the real-time collected videos, the test results of the traditional YOLOv3 and the proposed YOLOv3 utilization of road dense objects are shown as Figures 6 and 7, respectively.



(**a**) Close-range object detection (YOLOv3)

(**b**) Long-range object detection (YOLOv3)

Figure 6. Real-time video detection based on traditional YOLOv3.



(**a**) Close-range object detection (Proposed)

(**b**) Long-range object detection (Proposed)

Figure 7. Real-time video detection based on the proposed YOLOv3.

As seen by comparing Figures 6 and 7, the proposed YOLOv3 can detect and divide vehicles and pedestrians in real-time, and the detection confidence for different types of objects can also be provided.

Considering the difference in object image resolutions, and for the traditional YOLOv3, YOLOv5, YOLOv7 and the proposed YOLOv3, the detection time as shown in Tables 10 and 11 shows the corresponding detection time of selecting the length of 42.6 s of video as the input.

Table 10. Detection time for the images with different resolutions.

Algorithm	320 × 320 Detection Time/ms	416 × 416 Detection Time/ms	608 × 608 Detection Time/ms	FPS/Frames per Second
YOLOv3	22.8	29.4	51.9	43.6
YOLOv5	20.3	26.1	43.1	56.9
YOLOv7	19.8	26.7	41.2	59.7
The proposed YOLOv3	19.2	24.4	39.7	61.2

Table 11. Detection time for the video.

Algorithm	Detection Time/s	FPS/Frames per Second	FLOPS/Floating Point Operations per Second
YOLOv3	66.9	78	84.37 Bn
YOLOv5	64.7	85	86.38 Bn
YOLOv7	62.4	83	91.68 Bn
The proposed YOLOv3	19.2	84.4	98.7 Bn

Table 10 shows that the proposed YOLOv3 has the faster detection time, which changes proportionally with image resolution. Accordingly, the result can be taken from Table 10, which demonstrates that the proposed YOLOv3 can identify 61.2 frames per second, which is clearly superior to the 17.6 frames/s of the traditional YOLOv3 and superior to the 1.5 frames/s of the YOLOv7.

5. Conclusions

The application of the YOLOv3 network in road dense object detection is studied in this paper, mainly focusing on the deepening of Backbone network layers in the YOLOv3 network structure and the cross-layer addition and operation in the residual network. Different convolutional layers can realize the image detection of small, medium, and large scale features, respectively. Thus, the traditional idea of deepening and improving the recognition rate by relying on network structure is fundamentally improved. It can provide higher recognition accuracy with fewer network parameters and network layers, and the detection speed is also taken into account. The proposed algorithm can accurately distinguish different types of vehicles and pedestrians, even pedestrians at the edge of the detection area. A confidence box can be used to mark the probability of object classification, and the confidence does not decrease with the increasing of road object density. In contrast, although the YOLOv7 model has a good detection accuracy, the training and detection time is relatively low compared with the proposed algorithm, and it cannot be easily applied to real life. In addition, the proposed training strategy of transfer training in a special dataset can also be extended to other dense object detection scenarios, which can achieve high target detection precision and are simple to train. With the increasing investment of the government in transportation infrastructure and the continuous progress of science and technology, the technology of road vehicle testing is also constantly improving, and the accuracy and efficiency of testing are higher, which provides a good development opportunity for the road testing industry. At present, although the existing road detection algorithm can achieve a good detection accuracy, it cannot be applied to real scenarios on a large scale due to the large size of the network model and the slow detection time. Therefore, finding out how to improve the efficiency of the network model for road detection on the premise of ensuring detection accuracy should be the top priority for future research.

Author Contributions: Conceptualization, C.Z. and J.L.; Methodology, C.Z.; Software, J.L.; Validation, C.Z., J.L. and F.Z.; Formal analysis, J.L.; Investigation, J.L.; Resources, J.L.; Data curation, J.L.; Writing—original draft preparation, J.L.; Writing—review and editing, F.Z. and J.L.; visualization, J.L.; Supervision, C.Z. and F.Z.; Project administration, C.Z.; Funding acquisition, C.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was financially supported by the National Natural Science Foundation of China (61871176): Research of Abnormal Grain Conditions Detection using Radio Tomographic Imaging based on Channel State Information; Applied research plan of key scientific research projects in Henan colleges and Universities (22A510013): Research of Abnormal Grain Conditions Detection using Cone-beam CT based on Convolutional Neural Network; Open subject of Scientific research platform in Grain Information Processing Center (KFJJ2022011): Grain condition detection and classification-oriented semantic communication method in artificial intelligence of things; and The Innovative Funds Plan of Henan University of Technology Plan (2020ZKCJ02): Data-Driven Intelligent Monitoring and Traceability Technique for Grain Reserves.

Data Availability Statement: The results/data/figures in this manuscript have not been published elsewhere, nor are they under consideration (from you or one of your Contributing Authors) by another publisher.

Conflicts of Interest: All authors declare that they have no conflict of interest.

Nomenclature List

Abbreviation	Explanation
lbox	The error generated by the position in the prediction box
lobj	The error generated by confidence
lcls	The error generated by the class
i, j	Indicates the location of the predicted frame
x, y, w, h	Indicates the location of the prediction box

References

1. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
2. Oliveira, G.; Frazão, X.; Pimentel, A.; Ribeiro, B. Automatic graphic logo detection via Fast Region-based Convolutional Networks. In Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, 24–29 July 2016; pp. 985–991.
3. Chen, Y.; Li, W.; Sakaridis, C.; Dai, D.; Van Gool, L. Domain Adaptive Faster R-CNN for Object Detection in the Wild. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3339–3348.
4. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
5. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Part I 14, pp. 21–37.
6. Bell, S.; Zitnick, C.L.; Bala, K.; Girshick, R. Inside-Outside Net: Detecting Objects in Context with Skip Pooling and Recurrent Neural Networks. *Clin. Orthop. Relat. Res. (CoRR)* **2016**, 2874–2883.
7. Peng, Y.H.; Zheng, W.H.; Zhang, J.F. Deep learning-based on-road obstacle detection method. *J. Comput. Appl.* **2020**, *40*, 2428–2433.
8. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
9. Shao, L.; Zhu, F.; Li, X. Transfer learning for visual categorization: A survey. *IEEE Trans. Neural Netw. Learn. Syst.* **2014**, *26*, 1019–1034. [CrossRef] [PubMed]
10. Ai, H. Application of infrared measurement in traffic flow monitoring. *Infrared Technol.* **2008**, *30*, 201–204.
11. Zhao, Y.N. Design of a vehicle detector based on ultrasound wave. *Comput. Meas. Control.* **2011**, *19*, 2542–2544.
12. Jin, L. Vehicle detection based on visual and millimeter-wave radar. *Infrared Millim. Wave J.* **2014**, *33*, 465–471.
13. Wang, T.T. Application of the YOLOv4-based algorithm in vehicle detection. *J. Jilin Univ. (Inf. Sci. Ed.)* **2023**, *41*, 281–291.
14. Xiong, L.Y. Vehicle detection method based on the MobileVit lightweight network. *Appl. Res. Comput.* **2022**, *39*, 2545–2549.
15. Yuan, L. Improve the YOLOv5 road target detection method for complex environment. *J. Comput. Eng. Appl.* **2023**, *59*, 212–222.
16. Alshehri, A.; Owais, M.; Gyani, J.; Aljarbou, M.H.; Alsulamy, S. Residual Neural Networks for Origin–Destination Trip Matrix Estimation from Traffic Sensor Information. *Sustainability* **2023**, *15*, 9881. [CrossRef]
17. Xue, Q.W. Vehicle detection of driverless system based on multimodal feature fusion. *J. Guangxi Norm. Univ.-Nat. Sci. Ed.* **2022**, *40*, 6198.
18. Wang, W.H. Application research of obstacle detection technology in subway vehicles. *Comput. Meas. Control* **2022**, *30*, 110–116.
19. Huang, K.Y.; Chang, W.L. A neural network method for prediction of 2006 World Cup Football Game. In Proceedings of the 2010 International Joint Conference on Neural Networks (IJCNN), Barcelona, Spain, 18–23 July 2010; pp. 1–8.
20. Oyedotun, O.K.; El Rahman Shabayek, A.; Aouada, D.; Ottersten, B. Training very deep networks via residual learning with stochastic input shortcut connections. In Proceedings of the International Conference on Neural Information Processing, Guangzhou, China, 14–18 November 2017; Springer: Cham, Switzerland, 2017; pp. 23–33.
21. Zhu, B.; Huang, M.F.; Tan, D.K. Pedestrian Detection Method Based on Neural Network and Data Fusion. *Automot. Eng.* **2020**, *42*, 37–44.
22. Thomee, B.; Shamma, D.A.; Friedland, G.; Elizalde, B.; Ni, K.; Poland, D.; Borth, D.; Li, L.-J. The new data and new challenges in multimedia research. *Commun. ACM* **2015**, *59*, 64–73. [CrossRef]
23. Wang, S.; Huang, M.; Deng, Z. Densely connected CNN with multi-scale feature attention for text classification. In Proceedings of the IJCAI, Stockholm, Sweden, 13–19 July 2018; pp. 4468–4474.

24. Pan, S.J.; Kwok, J.T.; Yang, Q. Transfer learning via dimensionality reduction. In Proceedings of the AAAI, Chicago, IL, USA, 13–17 July 2008; Volume 8, pp. 677–682.
25. Rezende, E.; Ruppert, G.; Carvalho, T.; Ramos, F.; De Geus, P. Malicious software classification using transfer learning of resnet-50 deep neural network. In Proceedings of the 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), Cancun, Mexico, 18–21 December 2017; pp. 1011–1014.
26. Huang, K.; Chang, W. An Ground Cloud Image Recognition Method Using Alexnet Convolution Neural Network. *Chin. J. Electron Devices* **2010**, *43*, 1257–1261.
27. Cetinic, E.; Lipic, T.; Grgic, S. Fine-tuning convolutional neural networks for fine art classification. *Expert Syst. Appl.* **2018**, *114*, 107–118. [CrossRef]
28. Majee, A.; Agrawal, K.; Subramanian, A. Few-shot learning for road object detection. In Proceedings of the AAAI Workshop on Meta-Learning and MetaDL Challenge, PMLR, Virtual, 9 February 2021; pp. 115–126.
29. Xu, R.; Xiang, H.; Xia, X.; Han, X.; Li, J.; Ma, J. Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In Proceedings of the 2022 International Conference on Robotics and Automation (ICRA), IEEE, Philadelphia, PA, USA, 23–27 May 2022; pp. 2583–2589.
30. Dogo, E.M.; Afolabi, O.J.; Nwulu, N.I.; Twala, B.; Aigbavboa, C.O. A comparative analysis of gradient descent-based optimization algorithms on convolutional neural networks. In Proceedings of the 2018 international conference on computational techniques, electronics and mechanical systems (CTEMS), IEEE, Belgaum, India, 21–22 December 2018; pp. 92–99.
31. Zhang, K.; Xu, G.; Chen, L.; Tian, P.; Han, C.; Zhang, S.; Duan, N. Instance transfer subject-dependent strategy for motor imagery signal classification using deep convolutional neural networks. *Comput. Math. Methods Med.* **2020**, *2020*, 1683013. [CrossRef] [PubMed]
32. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162.
33. Wang, H.; Yu, Y.; Cai, Y.; Chen, X.; Chen, L.; Liu, Q. A comparative study of state-of-the-art deep learning algorithms for vehicle detection. *IEEE Intell. Transp. Syst. Mag.* **2019**, *11*, 82–95. [CrossRef]
34. Mao, H.; Yang, X.; Dally, W.J. A delay metric for video object detection: What average precision fails to tell. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 573–582.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.