Video Surveillance Vehicle Detection Method Incorporating Attention Mechanism and YOLOv5

Yi Pan*, Zhu Zhao, Yan Hu, Qing Wang
College of Intelligent Transportation, Hunan Communication Polytechnic, Changsha, China

***Abstract*—With the rising number of vehicle ownership nationwide and the consequent increase in traffic accidents, vehicle detection for traffic surveillance video is an effective method to reduce traffic accidents. However, existing video surveillance vehicle detection methods suffer from high computational load, low accuracy, and excessive reliance on large-scale computing servers. Therefore, the research will try to fuse coordinate attention mechanism to improve YOLOv5 network, choose lightweight YOLOv5s for image recognition, and use K-means algorithm to modify the aiming frame according to the characteristics of vehicle detection; meanwhile, in order to get more accurate results, coordinate attention mechanism algorithm, which is also a lightweight algorithm, is inserted into YOLOv5s for improvement, so that the designed The lightweight vehicle detection model can be run on embedded devices. The measurement experiments show that the YOLOv5+CA model completes convergence when the iterations exceed 100, and the localization loss and confidence loss gradually stabilize at 0.002 and 0.028, and the classification loss gradually stabilizes at 0.017. Comparing YOLOv5+CA with SSD algorithm, ResNet-101 algorithm and RefineDet algorithm, YOLOv5 +CA detection accuracy is better than other algorithms by about 9%, and the accuracy can be approximated to 1.0 at a confidence level of 0.946. The experimental results show that the research design provides higher accuracy and high computational efficiency for video surveillance vehicle detection, and can better provide reference value and reference methods for video surveillance vehicle detection and operation management.**

***Keywords*—Attention mechanism; YOLOv5; vehicle detection; image recognition; deep learning**

I. Introduction

After stepping into the 21st century, with the high-speed improvement of the economic level, vehicle ownership has been rising nationwide, and cars have become a common means of transportation, but at the same time, with the increase of vehicles, vehicle congestion, car accidents and other traffic problems are growing. At present, China's artificial intelligence technology continues to develop, automatic control technology tends to mature, combined with artificial intelligence and automatic control of intelligent traffic monitoring system has also been a large degree of development [1]. Intelligent Traffic System (ITS) can realize the organic integration of traffic system and various computer technologies, which can realize the instant, accurate and efficient management of traffic nationwide and effectively avoid a series of traffic congestion problems [2]. Among them, vehicle detection (VD) through video surveillance is the key to its capturing information, which can be applied to scenarios such as traffic flow calculation and violation vehicle capture. Vehicle target detection (TD) is a vital branch in computer vision. For the past few years, computer computing power integration is improving with the breakthrough of image acquisition equipment accuracy, this technology has received wide attention from researchers, while the breakthrough of algorithms in artificial intelligence (AI) has also benefited the field. VD through video surveillance joins artificial intelligence image recognition algorithms that mimic the human eye, which can sense and analyze targets in imitation of the human eye, and carry out the completion of vehicle recognition classification and localization [3]. Although the TD algorithm now has a high accuracy rate, but these functions need to rely on a powerful computing server, and in the daily VD its limited by the small volume of embedded equipment, cannot do large-scale computing. At the same time, when carrying out vehicle identification, due to changes in weather and lighting, there is a certain degree of difficulty for vehicle identification that blends into the background, and the identification accuracy is low in special environments such as rain and night. Based on this, to lift the accuracy of the video surveillance VD algorithm and solve the problem of excessive dependence of the algorithm on large computers, the study will try to combine attention mechanism neural network to improve detection accuracy using lightweight YOLOv5 network algorithm for research. Compared to similar literature, the study introduces the lightweight YOLOv5 algorithm to reduce the amount of computation in use and enable it to be loaded on small vehicles. The study also improves the lightweight YOLOv5 to improve the object recognition for subsequent use of the YOLOv5 recognition algorithm.

The study is divided into four parts. The first part provides an introduction to the integration of traffic systems with computer technology and the application of vehicle recognition therein, the second part discusses the related works in this domain, the third part uses an attention mechanism to improve the YOLOv5 algorithm to suit the vehicle recognition problem, the fourth part tests and analyses the performance of the model and algorithm; the fifth part concludes the above discussion.

II. Related Works

Vehicle detection (VD) is currently one of the main key-points of safety research in transportation, and the current situation of frequent traffic accidents has made experts aware of the value of the application. Deqing Liu et al. raised unmanned surface vehicle (USV) obstacle fusion detection based on Demperster-Shafer (D-S) evidence theory. The results show that multi-sensor fusion can use the complementarity among diverse sensors to supplement the obstacle detection details compared to the single-sensor detection method,

*Corresponding Author

effectively avoid the false detection of obstacles by single sensors, and show greater advantages in the reliability of obstacle detection [3]. Wang et al. aimed for improving the VD and tracking of autonomous vehicles using 3D Light Detection and Ranging (LiDAR) accuracy, a clustering algorithm trained by support vector machine (SVM) algorithm combined with Kalman filter and global nearest neighbor (GNN) algorithm is proposed to employ tracking of vehicles and further improve the accuracy of VD results with the help of tracking results [4]. Nguyen address the problem of large scale differences of vehicles and severe vehicle occlusion in VD by using a feature The results show better detection performance and lower computational cost [5]. Han et al. propose a CNN-M2R network with multilayer fusion and multidimensional attention to improve VD performance in urban areas, which uses a multidimensional attention network to highlight target convergence and a new difficulty-positive and negative sample balanced sampling strategy and a global balanced loss function to handle spatial imbalance and objective imbalance, the experimental results show a great improvement in detection performance compared to SSD, LRTDet, RFCN, and DFPN [6]. Saeed et al. focus on the often neglected last step of VD scheme deployment and design a single detector Mobile Net for embedded devices. A comprehensive deep-learning-based engineering VD solution is established and this solution has an average accuracy higher than 90% compared to common embedded devices, confirming the excellent real-time performance of the solution [7]. Liu et al. propose a backward feature enhancement network (BFEN) and a spatial layout preserving network (SLPN) to solve the interference caused by vehicle scale on VD in complex traffic scenarios and to accomplish accurate detection of miniature vehicles. Two-stage detector of SLPN is performed to achieve high recall detection of miniature vehicles. The method improves the competing baseline by 16.5% mAP, which has a better comparative performance compared to the current state of the art [8].

As an emerging intelligent network in the field of image recognition, YOLOv5 has been studied by a large number of scholars. Yan et al. proposed an intelligent classification method of coal gangue using YOLOv5 and multispectral imaging technology to deal with the issue of low accuracy and slow speed of traditional coal gangue recognition methods. The mean accuracy of gangue detection using YOLOv5.1 model reaches 98.34%, which could precisely identify gangue, as well as acquire gangue's relative position [9]. Jia et al. established a motorcycle helmet detection way combined with YOLOv5 for motorcycle driver helmet detection by video surveillance, which uses soft-NMS instead of NMS to fuse the YOLOv5 detector, and experimentally achieves 97.7% mAP, 92.7% F1 score and 63 frames per second (FPS), which is better than other methods [10]. Attention mechanism has also received a lot of attention after its introduction into artificial intelligence networks, and many scholars have conducted research on deep learning networks incorporating attention mechanism. Xu et al. developed a novel stock price prediction network on the basis of reinforcement learning (RL) through a bidirectional gated recurrent unit (GRU) network to better dig market changes from chaotic data for stock tendency. The model is superior to existing models and has excellent performance [11]. Lu et al. raised a 2-level interaction mode that relies on 2 time-varying attention mechanisms in order to accomplish the multi-person activity recognition task, and the model has high comparable performance, confirming the effectiveness of the attention mechanism [12].

In summary, although scholars have designed a large number of improved VD systems to improve the accuracy of video surveillance VD systems, there are still very few VD systems that have both high-speed computational effectiveness and lightweight embedded devices, both of which have strong potential applications in real-time VD.

III. Video Surveillance Vehicle Algorithm Design Based on YOLOv5 Network and Attention Mechanism

A. YOLOv5-based Video Surveillance VD Aiming Frame Improvement

The study was conducted to design algorithms aiming to accuracy lifting of VD through video surveillance, on the one hand, YOLOv5 (You Only Look Once fifth generation) was used as the baseline network, adding more techniques to improve the accuracy and speed, thus achieving a balance between accuracy and speed in the TD algorithm for vehicles, for another, attention mechanism was introduced to the VD algorithm for improvement to highly extract effective feature information highly relevant to VD and reduce the error brought by video surveillance. The improved algorithm for video surveillance measurement designed in the study is based on the YOLOv5 feature extraction network, which is OneStage series algorithm with the confidence level as in Equation (1).

$$Confidence = \Pr(Object) \times IOU_{pred}^{truth} \tag{1}$$

As shown in Equation (1), $\Pr(Object)$ denotes the possibility contained in the bounding box, which indicates the prediction accuracy adopting the loss function IOU, and the C conditional probability likelihood derivation performed in conjunction with this formula is Equation (2).

$$\Pr(Class_i / Object) \times \Pr(Object) \times IOU_{pred}^{truth} = \Pr(Class_i) \times IOU_{pred}^{truth} \tag{2}$$

In Equation (2), $\Pr(Class_i / Object)$ denotes the C conditional object probability in the grid, and the formula can indicate the matching degree in the prediction frame and object. In response to the high demand for timeliness of VD and the difficulty of the detection task, YOLOv5, which is the latest generation of algorithms, incorporates many techniques to improve accuracy and speed. It uses a loss function to evaluate the network effect, and the classification loss function is in Equation (2).

$$E_{cls} = \sum_{i=0}^{S^2} 1_i^{obj} \sum_{c \in classes} (p_i(c) - \hat{p}_i(c))^2 \tag{3}$$

In Equation (3), 1_i^{obj} denotes the objects in the grid i, $p_i(c)$ means the predicted possibility of the corresponding category, $\hat{p}_i(c)$ denotes the true probability, and S^2 denotes the number of grids into which the images are divided. The localization loss function is shown in Equation (4).

$$E_{box}=\lambda_{coord}\sum_{i=0}^{S^2}\sum_{j=0}^{B}1_{ij}^{obj}\left[\left(x_i-\hat{x}_i\right)^2+\left(y_i-\hat{y}_i\right)^2\right]+\lambda_{coord}\sum_{i=0}^{S^2}\sum_{j=0}^{B}1_{ij}^{obj}\left[\left(\sqrt{w_i}-\sqrt{\hat{w}_i}\right)^2+\left(\sqrt{h_i}-\sqrt{\hat{h}_i}\right)^2\right] \quad (4)$$

In Equation (4), λ_{coord} denotes the weight coefficient, x_i, y_i, w_i, and h_i are the prediction frame positions, $\hat{x}_i$, $\hat{y}_i$, $\hat{w}_i$, and $\hat{h}_i$ denote the true positions. B denotes the number of bounding boxes. The loss of confidence formula is Equation (5).

$$E_{obj}=\sum_{i=0}^{S^2}\sum_{j=0}^{B}1_{ij}^{obj}\left(C_i-\hat{C}_i\right)^2+\lambda_{noobj}\sum_{i=0}^{S^2}\sum_{j=0}^{B}1_{ij}^{noobj}\left(C_i-\hat{C}_i\right)^2 \quad (5)$$

In Equation (5), λ_{noobj} denotes the loss weight of objects not included in the bounding box, and C_i denotes the object i. The study selects the more lightweight YOLOv5s as the base-network for video surveillance VD, as shown in Fig. 1.

In Fig. 1, YOLOv5s' main division into input side, backbone, neck and head network is shown. The input data from the input side enters Focus to slice the picture, which extracts the pixel values in the picture every other value and slices a picture into four pictures in order to do improve the perceptual field and reduce the picture information loss. The above data enters the CSP layer after the convolution operation, which is an important concept of the YOLO series network. The formula of the convolution layer function used in this series of algorithms is shown in Equation (6).

$$a_j^l=f\left(b_j^l+\sum_{i\in M_j^l}a_j^{l-1}*k_{ij}^l\right) \quad (6)$$

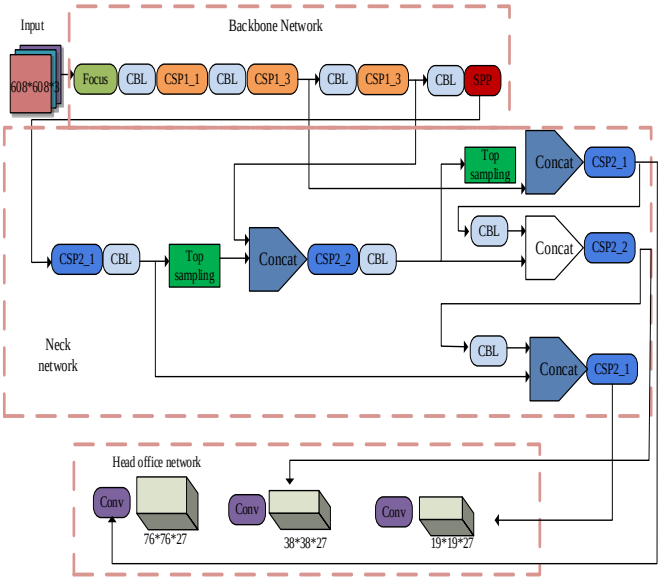

Fig. 1. YOLOv5 overall network structure.

As shown in Equation (6). In the convolutional layer l, a_j^l is the j-th output, b_j^l is the bias term, k_{ij}^l is the weight of the j-th convolutional kernel corresponding to the i-th position. a_j^{l-1} is the j-th output of the previous convolutional layer. f is the activation function. The computation of the feature map after the convolution operation is Equation (7).

$$out=\frac{in-k+2p}{s}+1 \quad (7)$$

In Equation (7), out means the output features size, in is the input graph size, k means the convolutional kernel size. s is the step size, and p denotes the width of the boundary fill. The CSP layer effectively avoids the problems of gradient information loss and network computation consumption during training of traditional large models, and effectively improves the learning capacity of the CNN, and its structure is shown in Fig. 2.

As shown in Fig. 2, the CSP structure of YOLOv5s divides the primordial input into two branches. After performing convolution operations, the amounts of channels are halved. Branch 1 performs Bottleneck*N, with two branches parallel, resulting in the same input and output sizes for bottleneck CSP. The CBL layer encapsulates three modules, namely BN, convolution layer and Leaky Relu activation function. BN is the original unit of the YOLO series, Equation (8).

$$\begin{cases}\hat{x}_i\leftarrow\dfrac{x_i-\mu_\beta}{\sqrt{\sigma_\beta^2+\varepsilon}}\\ y_i\leftarrow\gamma\hat{x}_i+\beta\end{cases} \quad (8)$$

As shown in Equation (8), μ_β and σ_β^2 denote the mean and variance of the data, $\hat{x}_i\leftarrow\dfrac{x_i-\mu_\beta}{\sqrt{\sigma_\beta^2+\varepsilon}}$ denotes the normalization of the sample, and $y_i\leftarrow\gamma\hat{x}_i+\beta$ denotes the translation and scaling of the data, with the BN function generally preceding the activation function.

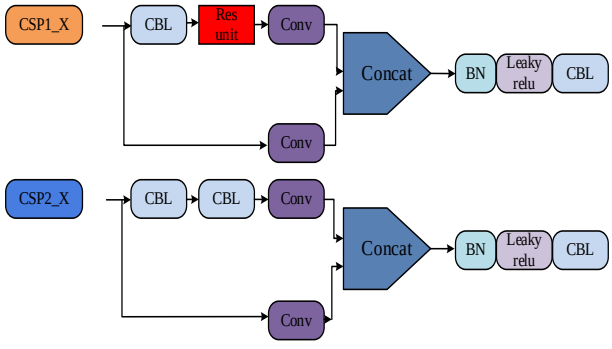

Fig. 2. CSP layer structure.

The CSP1 structure is mainly applied in Backbone, and that of CSP2 is mainly applied in Neck. The first parameter 1 in the CSP1_1 module indicates the CSP structure applied in Backbone Network, and the second parameter 1 indicates that the residual component in the module is repeated once. The CSP2x indicates the CSP module used in Neck network. CSP module. The main difference between it and the CSP module used in the Backbone Network is that 2X CBL modules are used instead of the residual module. Thereafter, the Neck network structure is entered by another convolution, which is schematically shown in Fig. 3.

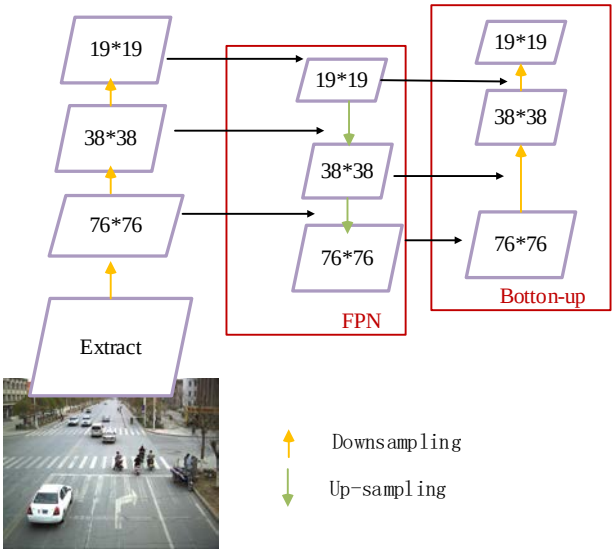

Fig. 3. Schematic graph of the neck network.

As Fig. 3, the extracted information is then input into the FPN module by performing a convolution operation on the target image extraction to reduce the image scale. The FPN module differs from the extraction module in that it passes feature information by a top-down approach, while the PAN module uses a bottom-up approach to pass feature information, which aims to enhance the target localization of network by using down sampling. The combination of these two structures not only enhances the target localization capability of the network, but also improves the target recognition capability, which helps to improve the accuracy of the TD algorithm.

The output is performed using the head network after passing through the neck network, and CIOU_Loss is used as the loss function of the Bounding box in Yolov5. The original algorithm of YOLOv5 is the result of the analysis of the COCO dataset, and the anchor box originally obtained from the COCO dataset setting is optimized in order to be more suitable for this VD experimental environment. The study uses the K-means algorithm for anchor selection frame optimization, which is essentially a clustering algorithm and belongs to the category of unsupervised learning. The error sum of squares is generally used as the objective function to categorize the samples, and this metric is often used to evaluate the effectiveness of the clustering results. Its specific expression is shown in Equation (9).

$$Loss = \sum_{i=1}^{k} \sum_{x \in c_i} dis(x, c_i) \tag{9}$$

As shown in Equation (9), *Loss* represents the error sum of squares, x represents the calculation sample, c_i represents the center of mass of the i category, and $dis(x, c_i)$ represents the distance between x and c_i. However, the formula is based on the Euclidean distance as an indicator for judging the similarity will cause more errors for big bounding boxes than for small bounding boxes, in order to make the K-means algorithm as an evaluation indicator for the similarity measurement of VD without the limitation of the bounding box size, the study uses A new distance formula suitable for VD, as in Equation (10).

$$dis(x, c_j) = 1 - IOU(x, c_j) \tag{10}$$

In Equation (10), x denotes the newly added checkbox, c_j denotes the first j real box, and $IOU(x, c_j)$ denotes the accuracy of the prediction of the location information of x and c_j using the loss function IoU (Intersection over Union). The improved K-means algorithm was tested by testing it on the UA_DETRAC dataset. The classification loss function is Equation (11).

$$E_{cls} = \sum_{i=0}^{S^2} 1_i^{obj} \sum_{c \in classes} \left(p_{i(c)} - \hat{p}(c)\right)^2 \tag{11}$$

As shown in Equation (11), $\sum_{i=0}^{S^2} 1_i^{obj}$ denotes the sum of squares over the objects in the table. $p_{i(c)}$ is the predicted rate of the corresponding category. $\hat{p}(c)$ is the true possibility.

B. Improvement of Video Surveillance VD Algorithm Based on Attention Mechanism

In the road information collected through video surveillance, there are not only target vehicles, but also contain invalid information such as pedestrians, trees, and barriers, which can interfere with VD, so the study uses attention mechanism to achieve target area locking to reduce the interference of invalid information.

Most attention mechanisms incorporated into neural network models provide some performance gains, but they are not as effective in lightweight networks as they are in large network models. Therefore, the study will use Coordinate Attention (CA) mechanisms that allow lightweight networks to obtain an extensive range of feature details and avoid introducing too much computational overhead, with the structure shown in Fig. 4.

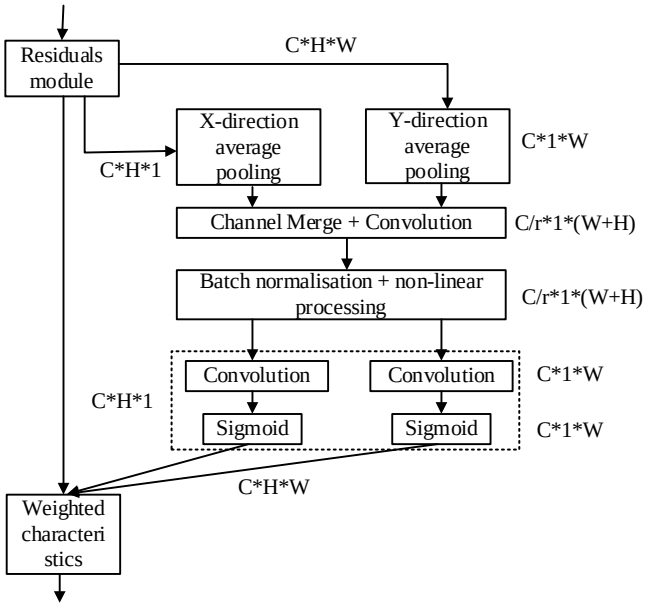

Fig. 4. Construction of the CA mechanism.

In Fig. 4, the CA mechanism is able to be identified as a computational unit to strengthen the characteristic representation capability of the mobile network, using two modules, coordinate information embedding and CA generation, to encode channel and long-distance relationships. Firstly, a 2D coordinate axis is created for the input feature information using a 1D global pooling operation, which is aggregated into 2 independent direction-aware feature representations along the X and Y directions. Then a merging operation is performed in the spatial dimension to integrate the feature maps using a 1*1 convolutional layer.

$$f(x)=\frac{1}{1+e^{-x}} \tag{12}$$

Finally, two attention weights are utilized to the input features by using a weighted multiplication of the Sigmoid function with normalized weights as in Equation (12), thus emphasizing the region of interest of the algorithm. The Sigmoid formula is as in Equation (12) and the tanh formula is as in Equation (13).

$$g(x)=\frac{1-e^{-2x}}{1+e^{-2x}} \tag{13}$$

Both Equation (12) and (13) use the exponential function e^{x} for the formulation. Applying the attention mechanism to the YOLOv5 can strengthen the recognition of target vehicles and the extraction of useful features for localization to a certain extent. The details are expressed in Table I.

TABLE I. COMPOSITION OF FEATURE EXTRACTION NETWORK WITH ATTENTION MECHANISM

Modules	Parameters			
	I	II	III	IV
Input	(640*640*3)	-	-	-
Focus	(3,64,1,1)	-	-	-
Conv	(64,128,3,2)	(128,256,3,2)	(256,512,3,2)	(512,1024,3,2)
3×C3	(128,128)	(1024,1024)	-	-
9xC3	(256,256)	(512,512)	-	-
SPP	(1024,1024,(5,9,13))	-	-	-
CoordAtt	(1024,1024)	-	-	-

Table I indicates the module name of the network structure, the first parameter in parentheses indicates the feature input channels of the mode, the second parameter indicates the feature output channel numbers, and the other parameters thereafter indicate the specific parameters of the module, for example, Focus (3,64,1,1), which indicates three input channels and 64 output channels, using a convolution of size 1*1. In introducing the coordinate attention mechanism into YOLOv5, the CA mechanism is first embedded into the backbone network of YOLOv5. Through existing research, it is found that in the YOLOv5 feature extraction network, the last layer has the largest number of feature channels, which may affect the accuracy of the detection algorithm due to the interference of irrelevant information, so the model is added to the last layer in an attempt to allow the VD algorithm can focus on the feature information related to the current task.

The evaluation metrics of the YOLOv5 algorithm improved by the fused attention mechanism are selected as accuracy, recall and detection speed, and the evaluation metrics are calculated using the confusion matrix as the basis, and the accuracy (Precision) is denoted by P. In the confusion matrix, it indicates what percentage of the results with positive prediction are predicted correctly, as in Equation (14); in the confusion matrix, TP is both the predicted and true cases are active cases. FP is that the prediction is positive and the true case is inactive.

$$\Pr ecision=\frac{TP}{TP+FP} \tag{14}$$

Recall (Recall) in the confusion matrix indicates what percentage of all positive columns are predicted, as in Equation (15).

$$\mathrm{Re}\,call=\frac{TP}{TP+FN} \tag{15}$$

FN is the opposite of *FP* . In the current experimental algorithm it is not possible to obtain results with high recall and accuracy, so the equilibrium state of the two needs to be considered [13]. Based on the derived recall and accuracy, the average precision (AP) is considered, and since this experiment uses a multi-objective VD algorithm, it is measured using the category-wide average precision metric, which is obtained by weighting the mean precision of all detection categories [14]. At the same time, the study is a lightweight model, and for achieving the effect of saving computational materials, it is also necessary to evaluate the detection speed, and Frame Per Second (FPS) is selected as the speed evaluation index.

IV. Performance Testing of YOLOv5 Built on Improved Attention Mechanism

A. Experimental Scheme Design and Computational Efficiency Analysis

For testing the recommendation model, a test experiment is designed here. In order to meet the requirements of diverse road conditions and real scene fit, the study will use the UA_DETRAC dataset, which is obtained by slicing the traffic routes of Chinese cities Beijing and Shanghai at 25 frames per second after 10 hours of video shooting, with the image size of 960*540 pixels, containing more than 140,000 images, and manually labeled 8250 vehicles, with 1.21 million bounding boxes marked, and car types classified according to vehicle shape, and all marked vehicles are dynamic vehicles, and the shooting environment includes four kinds: sunny day, rainy day, cloudy day and night shown in Fig. 5.

Fig. 5. Sample diagram of part of the UA_DETRAC dataset.

Firstly, to reduce the burden on the experimental equipment and remove the data redundancy of the dataset, UA_DETRAC was extracted into a new dataset with a ratio of 5:1 and converted into a dataset in VOC format, and the annotation file format was converted from xml to txt for easy input into the YOLOv5 model for model training. Because the size of UA_DETRAC images is 960*540 pixels, the images are scaled to the same size of 640*640 pixels. After training the model, the α-CloU loss function curve localization loss curve, classification loss curve and confidence loss curve are shown in Fig. 6.

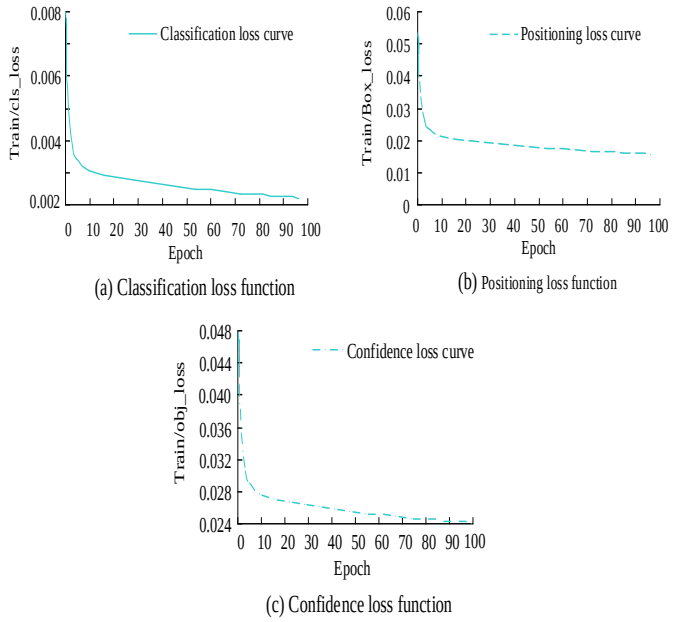

Fig. 6. YOLOv5 model training loss function curve.

The changes of loss function curves during training are Fig. 6. Fig. 6(a) indicates the classification loss curve. Fig. 6(b) shows the localization (positioning) loss curve, and Fig. 6(c) indicates the confidence loss curve. During the training process, no abnormalities have occurred, and all the loss function curves tend to be stable when the model is trained to the 100th round. From Fig. 6, it can be seen that the localization loss and confidence loss gradually stabilize at 0.002 and 0.028; the classification loss gradually stabilizes at 0.017.

Based on this, the algorithm is tuned to improve the focus on the key regions by adding a coordinate attention mechanism to the YOLOv5 model that mimics human visual recognition and has lightweight characteristics. Before training, the hyperparameter batch size is 16, and 100 epochs are trained. From the beginning to the end of training, the warm-up principle is used, which means that 3 epochs are learned from 0. After the learning rate reaches a plateau, the cosine annealing principle is adopted to reduce the learning rate, and the cosine annealing hyperparameter is set to 0.2 [15]. For the selection of the optimizer, the study Random Gradient Descent with momentum was chosen. The advantage of this method is that the square of the gradient is calculated in a small space, so there is no need to store the gradient. The momentum of the optimizer is 0.937 and the weight decay coefficient is 0.0005. After turning on mosaic data enhancement for all training images, mix up data enhancement is turned off. The loss function curve of the training result of the network model with the CA mechanism added is shown in Fig. 7.

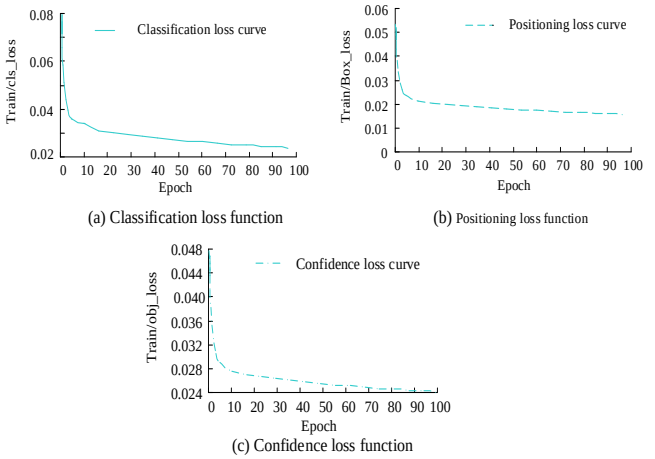

Fig. 7. Change curve of the loss function of the YOLOv5 model with the introduction of the coordinate attention mechanism.

As shown in Fig. 7, 7(a) represents the classification loss curve, Fig. 7(b) represents the localization loss curve, and Fig. 7(c) represents the confidence loss curve. The overall training process of the model is relatively normal, with a smooth decreasing trend, and the loss decreases quicker in the 1st 20-epochs, and then gradually stabilizes when reaching 100 epochs. The final localization loss is stabilized at 0.017, classification loss is stabilized at 0.00117, and confidence loss is stabilized at 0.028.

B. Model Vehicle Inspection Quality Analysis

For verifying the effectiveness of CA mechanism on the improvement of YOLOv5 measurement accuracy, Squeeze and Excitation Networks (SE), Convolutional Block Attention Module (CBAM) and CA mechanism are added to the YOLOv5 model, respectively. The SE is added at the same location as the CA mechanism, and the CBAM has the ability to extract spatial information instead of the convolutional layer, so the CBAM is used to replace that in the 5th-layer of the YOLOv5 [16-18]. The common YOLOv5 model is also selected as a comparison, and the accuracy comparison curves of the four models are Fig. 8.

As shown in Fig. 8, Fig. (a), (b), (c), and (d) show the test results of the baseline network models YOLOv5, YOLOv5+SE, YOLOv5+CBAM, and YOLOv5+CA, respectively, and it can be seen that the overall trend of the four network models is similar, and all of them gradually stabilize after a rapid increase in the confidence interval from 0 to 1. Further analysis of Fig. 8(a) and Fig. 8(b) shows that the accuracy of both models can be approximated to 1.0 at a confidence level of 0.946 for all categories, but the accuracy curve of YOLOv5 model with SE inserted for other categories of vehicles is smoother and has higher validity than the accuracy curve of YOLOv5 baseline model. In comparing Fig. 8(c) with Fig. 8(d), the YOLOv5 model with CA inserted has a higher confidence level of correct prediction for all categories of VD, and has a good accuracy at a confidence level of 0.4. The comparison of the accuracy curves demonstrates that the fused CA mechanism YOLOv5 network model performs better in VD. After that, the average accuracy of the four models is compared (Fig. 9).

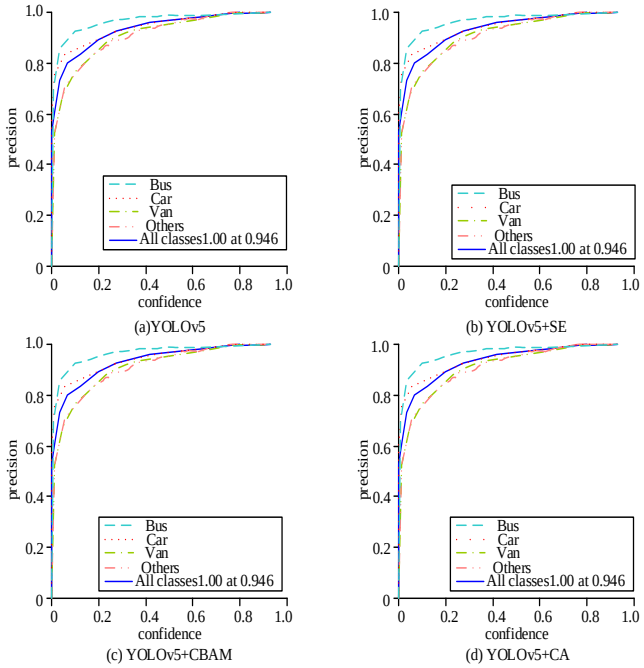

Fig. 8. Comparison of the precision of the improved model with the baseline network.

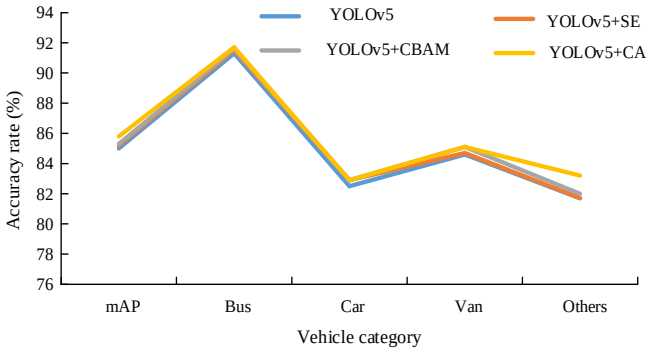

Fig. 9. Comparison of detection accuracy of four improved VD algorithms.

As shown in Fig. 9, which clearly reflects the comparison of the detection accuracy of the four VD algorithms network, to better reflect the improvement of the detection accuracy of the VD algorithms, also the more common VD algorithms and YOLOv5 combined with the CA mechanism of the network for comparison. It can be clearly found that the overall trend of the four VD algorithms is similar, and the detection accuracy in bus classification is much higher than the other classifications, at more than 90%, and the detection accuracy of the four algorithms for car classification is lower all approximating 83%, which may be due to the relatively fixed bus shape with obvious signs [19-21]. The YOLOv5 performance combined with CA mechanism for VD algorithm is significantly greater than the others, with detection accuracy exceeding other algorithms by about 0.8 percentage points.

As shown in Fig. 10, it more intuitively demonstrates the improvement of YOLOv5+CA on VD accuracy, a comparison using the commonly used VD algorithm model SOTA [22], it is evident that the YOLOv5+CA VD algorithm performs significantly better than several other common algorithms, with

detection accuracy exceeding other algorithms by about 9 percentage points. Unlike the YOLOv5+CA detection algorithm, the SSD algorithm, ResNet-101 algorithm and RefineDet algorithm have similar trends, and the YOLOv5+CA detection algorithm is significantly more accurate than the three common algorithms in detecting mAP, Bus, Van and other categories, while the three common algorithms perform slightly better in detecting vehicles in the Car category in terms of accuracy than the YOLOv5+CA detection algorithm.

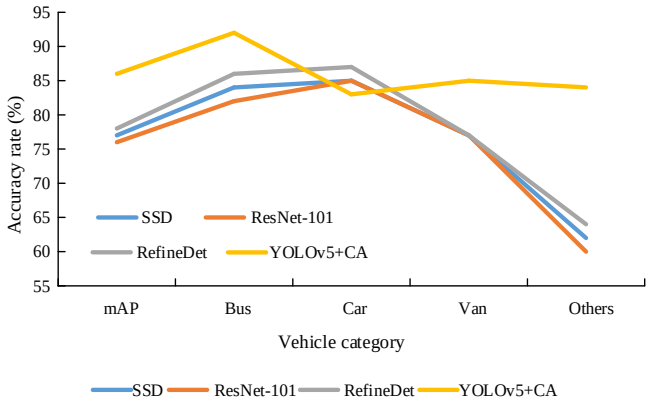

Fig. 10. Accuracy comparison of YOLOv5+CA with different SOTA vehicle detection algorithms.

As shown in Fig. 11, the YOLOv5 combined with the three attention mechanisms is tested using night-time live video, and the detection results are visualized and compared. Fig. 11(a), 11(b) and 11(c) show YOLOv5+SE, YOLOv5+CBMA and YOLOv5+CA respectively by comparing Fig. (a), (b) and (c) it can be found that three small cars are missed in the figure. In Fig. 11(b) the VD using the network of YOLOv5 combined with CBAM leads to an improvement of the missed detection, two of the missed vehicles are identified, but the confidence level is not high. In Fig. 11(c), the algorithm using YOLOv5+CA identifies all three missed vehicles with high confidence values. Therefore, the validity of the CA mechanism in improving the accuracy of the VD algorithm can be fully confirmed.

Fig. 11. Visual comparison of three YOLOv5 test results.

V. Conclusion

For the problem of low accuracy of VD performed by video surveillance, an improved VD model incorporating coordinate attention mechanism and YOLOv5 is designed in the study. The experimental results of the performance test show that the model converges when the iterations exceed 100, and the localization loss and confidence loss gradually stabilize at 0.002 and 0.028; the classification loss gradually stabilizes at 0.017. After adding the CA mechanism and setting the training to 100 epochs, the loss of the model declined quicker in the 1st 20 epochs. After that, the loss leveled off when reaching 100 epochs. Finally, the localization loss is stabilized at 0.017, the classification loss is stabilized at 0.00117, and the confidence loss is stabilized at 0.028. Thereafter, to compare the superiority of different attention mechanisms, the baseline network models YOLOv5, YOLOv5+SE, YOLOv5+CBAM, and YOLOv5+CA are used for comparison tests, and the results show that the accuracy can be approximated to 1.0 at a confidence level of 0.946 for all categories, and the YOLOv5 model with CA inserted for all categories of VD predicts the correct the confidence level is higher and has a good accuracy at a confidence level of 0.4. To better reflect the improvement in detection accuracy of the VD algorithm, YOLOv5+CA is compared with SSD algorithm, ResNet-101 algorithm and RefineDet algorithm, and the results show that YOLOv5+CA VD algorithm performs significantly better than several other common algorithms, and the detection accuracy is better than other algorithms by about 9%. To compare the actual video VD gap of YOLOv5+SE, YOLOv5+CBAM, and YOLOv5+CA, a visual comparison of YOLOv5 combining the three attention mechanisms reveals that both YOLOv5+SE and YOLOv5+CBAM have missed detections and low confidence levels. In summary, the research has shown that the vehicle detection accuracy of the YOLOv5+CA model designed by the research is higher than that of common models, and the computational performance has been improved compared to the original algorithm, but there are still some shortcomings, and the subsequent research can be improved and improved from the following aspects: (1) producing vehicle datasets with higher image quality. When using deep learning methods, the quality of the dataset determines the upper limit of the algorithm's performance. (2) Improve the detection accuracy of small targets. (3) Further compression of the network model. Due to the limited computing power and storage space of actual embedded devices, large scale deep learning algorithms cannot yet be deployed into these devices. And deploying the algorithms in embedded devices can reduce the latency time due to data passing through the network. The current lightweighting tends to sacrifice a certain amount of accuracy, and how to make the algorithm perform model compression while keeping accuracy constant is also a direction for future research. (4) Later on, consideration can also be given to improving the robustness of the algorithm to cope with different weather conditions.

References

[1] Islam N, Phillips C. Intelligent Traffic Engineering, TE) system for rural broadband. Computer networks, 208, May 8, 1088-1100, 2022.

[2] Hu R, Xu Y, Chen H, Zou F. A novel method for the detection of road intersections and traffic rules using big floating car data. IET intelligent transport systems, 16, 8, 983-997, 2022.

[3] Liu D, Zhang J, Jin J, Dai Y, Li L. A new approach of obstacle fusion detection for unmanned surface vehicle using Dempster-Shafer evidence theory. Applied Ocean Research, 119, 4, 103-116, 2022.

[4] Wang H, Zhang X. Real-time vehicle detection and tracking using 3D LiDAR. Asian Journal of Control: Affiliated with ACPA, the Asian Control Professors Association, 24, 3, 1459-1469, 2022.

[5] Nguyen H. Multiscale feature learning based on enhanced feature pyramid for vehicle detection. Complexity, 2021, 20, 121-131, 2022.

[6] Han Z, Wang C, Fu Q. M~2R-Net: deep network for arbitrary oriented vehicle detection in MiniSAR images. Engineering Computations: International Journal for Computer-Aided Engineering and Software, 38, 7, 2969-2995, 2022.

[7] Saeed A, Haghighat A, Sharma A. A deep-learning-based computer vision solution for construction vehicle detection. Computer-Aided Civil and Infrastructure Engineering, 35, 7, 753-767, 2022.

[8] Liu W, Liao S, Hu W. Towards accurate tiny vehicle detection in complex scenes. Neurocomputing, 347, Jun. 28, 24-33, 2019.

[9] Yan P, Sun Q, Yin N, Hua L, Shang S, Zhang C. Detection of coal and gangue based on improved YOLOv5.1 which embedded scSE module*. Measurement, 26, 7, 530-542, 2022.

[10] Jia W, Xu S, Liang Z, Zhao Y, Min H, Li S, Yu Y. Real-time automatic helmet detection of motorcyclists in urban Real-time automatic helmet detection of motorcyclists in urban traffic using improved YOLOv5 detector. IET Image Processing, 15, 14, 3623-3637, 2021.

[11] Xu H, Chai L, Luo Z, Li S. Stock movement prediction via gated recurrent unit network based on reinforcement learning with incorporated attention mechanisms. Neurocomputing, 467, Jan. 7, 214-228, 2022.

[12] Lu L, Di H, Lu Y, Zhang L, Wang S. A two-level attention-based interaction model for multi-person activity recognition. Neurocomputing, 322, Dec. 17, 195-205, 2018.

[13] Barma M, Modibbo U M. Multiobjective mathematical optimization model for municipal solid waste management with economic analysis of reuse. Journal of Computational and Cognitive Engineering, 1, 3, 122-137, 2022.

[14] Voskoglou M G. A combined use of soft sets and grey numbers in decision making. Journal of Computational and Cognitive Engineering, 2, 1, 1-4, 2023.

[15] Maihulla A S, Yusuf I, Bala S I. Reliability and performance analysis of a series-parallel system using Gumbel-Hougaard family copula. Journal of Computational and Cognitive Engineering, 1, 2, 74-82, 2022.

[16] Zeeshan Z, Ain Q U, Bhatti U A, Memon W H, Ali S, Nawaz S A, Nizamani M M, Mehmood A, Bhatti M A, Shoukat M U. Feature-based multi-criteria recommendation system using a weighted approach with ranking correlation. Intelligent Data Analysis, 25, 4, 1013-1029, 2021.

[17] Salina A, Ilavarasan E, Rao K Y. IoT enabled machine learning framework for social media content based recommendation system. International Journal of Vehicle Information and Communication Systems, 7, 2, 161-175, 2021.

[18] Bhuvaneshwari P, Rao A N. Product recommendation system using optimal switching hybrid algorithm. International Journal of Intelligent Enterprise, 8, 2/3, 185-204, 2021.

[19] Sundari P S, Subaji M. A comparative study to recognize fake ratings in recommendation system using classification techniques. Intelligent Decision Technologies: An International Journal, 15, 3, 443-450, 2021.

[20] Cui Y. Intelligent recommendation system based on mathematical modeling in personalized data mining. Mathematical Problems in Engineering, 2021, 3, 2036-2047, 2021.

[21] Azimirad V, Sani M F. Experimental study of reinforcement learning in mobile robots through spiking architecture of Thalamo-cortico-thalamic circuitry of mammalian brain. Robotica, 38, 9, 1558-1575, 2021.

[22] Saraswathi K, Mohanraj V, Suresh Y, Senthilkumar J. A hybrid multi-feature semantic similarity based online social recommendation system using CNN. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems: IJUFKS, 29, Dec. Suppl. 2, 333-352, 2021.