Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.056241

ARTICLE

Special Vehicle Target Detection and Tracking Based on Virtual Simulation Environment and YOLOv5-Block+DeepSort Algorithm

Mingyuan Zhai[1,2], Hanquan Zhang[1], Le Wang[1], Dong Xiao[1,*], Zhengmin Gu[3] and Zhenni Li[1]

[1]College of Information Science and Engineering, Northeastern University, Shenyang, 110819, China

[2]Flight Control Department, Shenyang Aircraft Design and Research Institute, Shenyang, 110035, China

[3]Department of Information Center, The First Hospital of China Medical University, Shenyang, 110001, China

*Corresponding Author: Dong Xiao. Email: xiaodong@ise.neu.edu.cn

Received: 17 July 2024 Accepted: 24 September 2024 Published: 18 November 2024

ABSTRACT

In the process of dense vehicles traveling fast, there will be mutual occlusion between vehicles, which will lead to the problem of deterioration of the tracking effect of different vehicles, so this paper proposes a research method of virtual simulation video vehicle target tracking based on you only look once (YOLO)v5s and deep simple online and realtime tracking (DeepSort). Given that the DeepSort algorithm is currently the most effective tracking method, this paper merges the YOLOv5 algorithm with the DeepSort algorithm. Then it adds the efficient channel attention networks (ECA-Net) focusing mechanism at the back for the cross-stage partial bottleneck with 3 convolutions (C3) modules about the YOLOv5 backbone network and before the up-sampling of the Neck feature pyramid. The YOLOv5 algorithm adopts expected intersection over union (EIOU) instead of complete intersection over union (CIOU) as the loss function of the target frame regression. The improved YOLOv5 algorithm is named YOLOv5-Block. The experimental results show that in the special vehicle target detection (TD) and tracking in the virtual simulation environment, The YOLOv5-Block algorithm has an average accuracy (AP) of 99.5%, which significantly improves the target recognition correctness for typical occlusion cases, and is 1.48 times better than the baseline algorithm. After the virtual simulation video sequence test, multiple objects tracking accuracy (MOTA) and various objects tracking precision (MOTP) improved by 10.7 and 1.75 percentage points, respectively, and the number of vehicle target identity document (ID) switches decreased. Compared with recent mainstream vehicle detection and tracking models, the YOLOv5-Block+Deepsort algorithm can accurately and continuously complete the detection and tracking tasks of special vehicle targets in different scenes.

KEYWORDS

YOLOv5; DeepSort; virtual simulation; attention mechanism

1 Introduction

With the development of computer vision, deep learning (DL) methods are widely used in TD [1]. Moving TD and tracking is an extremely important exploration content in the field of computer vision. It is also a key part of artificial intelligence technology. DL-based TD and tracking algorithms (TA) have their unique advantages, with higher detection accuracy in various complex environments,

support for simultaneous detection of multiple targets, greater scalability, and better robustness against interference factors [2–4]. It will speed up the review and typesetting process. It has become a research hotspot in the fields concerning intelligent security, driverless driving, and mobile robots, and has shown very broad application prospects. With the rapid improvement of the high-intensity and fast-paced global informatization level, the detection regarding vehicle targets puts forward higher requirements and challenges for manpower and material resources, and it is difficult to guarantee the real-time and accuracy of the detection results regarding input images [5,6]. Currently, the detection and tracking technology for small and medium-sized vehicles has been relatively mature. However, the detection and tracking of special vehicles has the problem that the vehicles are large and easily block each other. This situation will reduce the detection accuracy [7].

Effective detection of vehicle targets is the basis concerning tracking. Traditional target detection algorithms are mainly based on manual feature extraction. First, the region of interest is selected, the candidate regions are obtained by using sliding windows to exhaustively traverse the entire image, then all these candidate regions are extracted with features, and finally, the classifier is used for recognition. The current mainstream TD algorithms are mainly based on DL algorithms, mainly divided into two-stage TD and single-stage TD. In 2014, Girshick et al. [8] proposed the region convolutional neural networks (R-CNN) algorithm, which substantially improved the accuracy of TD, but the computational efficiency was low because R-CNN was extracting and computing features for all candidate regions. In the same year, He et al. [9] proposed spatial pyramid pooling networks (SPP-Net) which effectively solved the problem regarding repeated computation for R-CNN, but the problem concerning excessive memory occupation about R-CNN was not solved. In 2015, fast R-CNN [10], which was built on SPP-Net, improved the computation speed and solved the problem of excessive memory occupation at the same time. Faster R-CNN [11] proposed to use of RPN (Region Proposal Network) for edge training for the problem regarding selective search, which effectively improved the detection speed. In 2016, the YOLO [12] algorithm attracted a lot of attention for its excellent speed. In the same year, a single shot detector (SSD) [13] algorithm proposed a method to detect targets in images using a single deep neural network. It combines the advantages of Faster R-CNN and YOLO to improve the detection speed while ensuring detection accuracy. Subsequently, the YOLO algorithm is continuously optimized in continuous iterations. Based on YOLOv1, YOLOv2 [14] solves the problems concerning overfitting and unreasonable pre-set prior boxes and further improves the detection speed and accuracy. Then, YOLOv3 [15] was comprehensively upgraded based on YOLOv2, especially strengthening the detection capability of small targets while maintaining an efficient detection speed. In April 2020, the author concerning YOLOv4 [16] added spatial pyramid pooling (SPP) and parallel attention network (PANet) [17] to the CSP network developed on YOLOv3, further improving the detection speed and accuracy. Two months later, YOLOv5 opened its source code on Git Hub, further accelerating the speed while improving the accuracy, which is conducive to the deployment concerning mobile terminals, and it is the most used TD algorithm in the industry.

In multi-target tracking for special vehicles, the main implementation is to continuously localize multiple targets in consecutive video frames, maintain the ID concerning each target, and record the trajectory information [18–20]. The multi-target tracking is built upon the foundation of single-target tracking, where the single-target tracking algorithm initially acquires a singular target position in the initial frame and subsequently predicts the size and position of that target continuously across subsequent frames. Typical algorithms include the online learning-based long-time TA tracking-learning-detection [21], the Kalman filter-based algorithm [22], the kernel correlation filter-based algorithm kernelized correlation filters (KCF) [23], etc. The multi-target algorithm is more complex and needs to deal with the appearance and disappearance of targets, motion prediction about tracking

targets, target matching between upper and lower frames, and occlusion and scale transformation processing of tracking targets. Bewley et al. [24] proposed an online real-time multi-target TA, simple online and real-time tracking (SORT), which uses a combination of Kalman filtering and the Hungarian correlation algorithm. The algorithm greatly improves the tracking performance and speed by optimizing the detector. Later, Wojke et al. [25] proposed the DeepSort algorithm, which uses the intersection over union (IOU) fusion metric to calculate the degree of match between TD and target tracking framework, improving the ID jumping problem caused by occlusion and other factors during multi-target tracking, and making the performance about multi-target tracker. The performance concerning multi-target tracker rises to a new level. Tian et al. [26] proposed a shallow feature fusion algorithm based on SORT, shallow feature fusion algorithm based on SORT (SFFSORT), which has better experimental results in tracking and detecting urban road traffic than SORT and the original DeepSort, but it has not been applied to the traffic scene for special vehicles.

Vehicle tracking based on virtual simulation video is a typical multi-target tracking task. The difficulties lie in occlusion between vehicles, motion blur, scale changes, etc. Li et al. [27] proposed a method based on the combination concerning YOLOv3 and KCF in the vehicle motion trajectory extraction task. This method associates the TD results with the prediction results of historical trajectories to complete target vehicle tracking. However, the KCF algorithm is not suitable for target scale changes and rapid deformation and is not suitable for vehicle tracking in surveillance video scenes. To solve the problems of slow recognition speed and low statistical accuracy in traffic flow statistics, Liu et al. [28] proposed YOLO recognition and Mean shift tracking methods. The Mean shift method can update the target model based on the recognition results of YOLO to improve the problem of tracking failure. However, the Mean shift algorithm only uses a single-color feature to describe the target and is not robust enough to interfere with similar color targets. Jin et al. [29] proposed a forward multi-vehicle target TA optimized for DeepSort. This method uses improved YOLOv3 as the detector and performs re-identification pre-training on the amplified VeRi dataset, combining the center loss function and the cross-entropy loss function. Compared with the original DeepSort, the experimental results have improved accuracy and the number of IDs has decreased. Bin Zuraimi et al. [30] applies the original YOLOv4 algorithm to realize vehicle detection on public transportation roads and also uses the DeepSort algorithm to calculate the number of passing vehicles in a specific video. However, YOLOv4 and DeepSort algorithms are used separately, and the two are not combined. Zhao et al. [31] studied multi-vehicle tracking. The method uses lightweight and efficient YOLOv5s, adding convolutional block attention module (CBAM) and Transformer encoder modules to provide high-confidence detection and trajectories for subsequent data association, which can reduce identity switching due to mutual occlusion.

On this basis, this paper proposes and applies a detection algorithm for special vehicles (such as tanks and armored vehicles, etc.) based on the virtual simulation scenario, which can accurately identify the blocked vehicle targets. In addition, this paper also studies the special vehicle tracking algorithm based on DeepSort. Combining the YOLOv5-Block proposed in this paper with DeepSort, it can stably track moving special vehicle targets in virtual simulation videos. The main contributions of this paper are as follows:

1) In this paper, a new large vehicle detection model YOLOv5-Block algorithm is introduced. Vehicle target features are highlighted by incorporating the ECA-Net attention mechanism in the C3 module regarding the YOLOv5s backbone. Secondly, after multiple convolution operations, a lot of important information will be lost. The ECA-Net attention mechanism is added before up-sampling on the Neck feature pyramid, which strengthens the backbone's

ability to detect dense vehicle targets and significantly improves the detection accuracy for special vehicles.

2) Proposed a new large-vehicle detection and tracking framework based on the combination of YOLOv5-Block and DeepSort algorithm. The framework effectively reduces the difficulty of vehicle identity switching caused by occlusion, and thus effectively detects and tracks the accuracy rate.
3) EIOU is used instead for CIOU as the loss function about target box regression in the YOLOv5-Block algorithm, which makes the target box regression process more stable and the convergence accuracy higher. The improved YOLOv5-Block+DeepSort algorithm can effectively improve the omission, misdetection, and ID switch that may be caused by occlusion and insufficient light in the virtual simulation scene.
4) Since there are no publicly available image datasets concerning special vehicles (such as tanks and other armored vehicles) in battlefield environments, we used a simulation platform to build a large, high-quality dataset of 3886 images about a variety of special vehicles. The dataset can be used for special vehicle detection and tracking tasks while solving the problem concerning the unavailability of public datasets.

The rest of this article is organized as follows. We first provide a methodology for our work in Section 2, describing our approach to vehicle detection and tracking. Section 3 details the specific improvements made. Section 4 is the experiment and result analysis. Finally, Section 5 is the summary of this paper.

2 Methodology

2.1 YOLOv5 Structure

The information adaptive method of YOLOv5 is: firstly, calculate the scaling concerning the original image according to the scaling size, for different sizes concerning the original image will get a variety of scaling factors, choose the smallest of them; secondly, multiply both the length and width regarding original image by the scaling factor, and then fill it to the standard scaling size. In this paper, the image size of input YOLOv5 is set to 608×608.

The structure of YOLOv5 includes four parts: Input, Backbones, Neck, and Prediction, as shown in Fig. 1. In the Input stage, YOLOv5 preprocessed the input data through mosaic data enhancement, dynamic anchor frame mechanism, and adaptive image adjustment technology to improve the generalization ability about the model. The Backbones section integrates cross-stage partial (CSP) and spatial pyramid pooling (SPPF) technologies. The CSP structure is designed to reduce the amount concerning computation and improve the efficiency of model, while the SPPF enhances the detection accuracy for model through multi-scale pooling operation. The Neck part uses an architecture that combines feature pyramid network (FPN) and PANet, where down-sampling operations help to enhance semantic information extraction while up-sampling helps to pinpoint information. Finally, the Prediction section serves as the output of the entire model and is responsible for the final detection result.

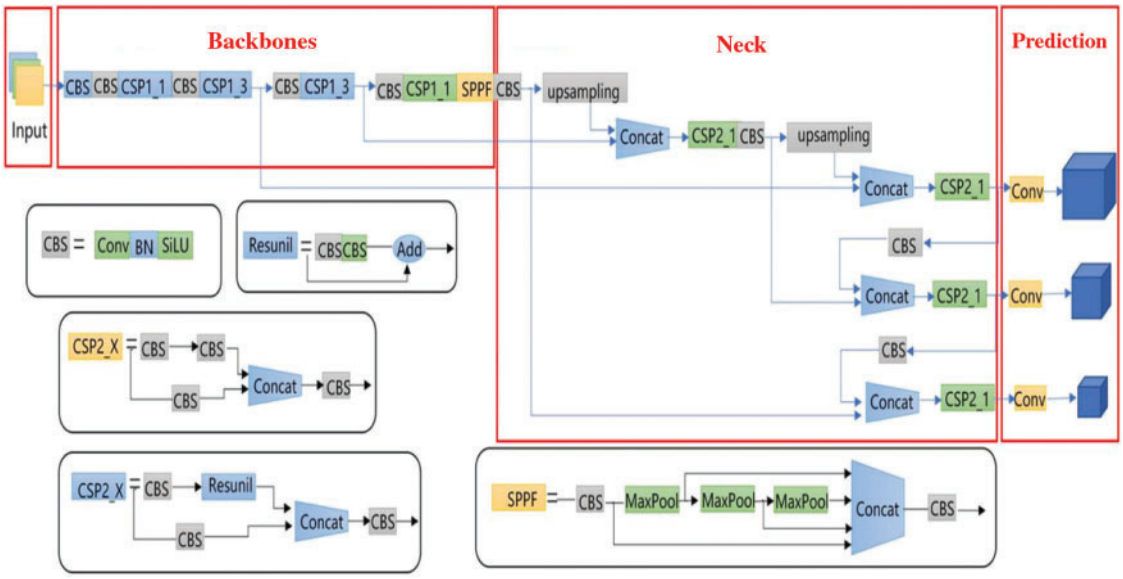

Figure 1: YOLOv5 structure

2.2 DeepSort Algorithm

This study uses DeepSort as the TA. The DeepSort algorithm incorporates a re-identification model to augment the matching efficacy of the Hungarian algorithm through the integration of appearance and motion information, thereby mitigating ID-switching (IDSW) occurrences. The overall process regarding DeepSort algorithm is to predict the input motion trajectory for the TD algorithm using the Kalman filter, and then use the Hungarian algorithm to match the predicted trajectory with the current frame detection result in cascade and IOU matching, and finally repeat the above steps to obtain the complete tracking trajectory. DeepSort flowchart is shown in Fig. 2.

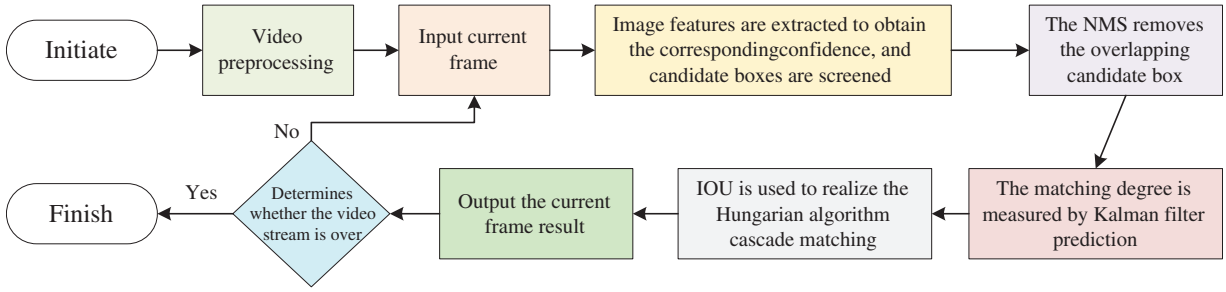

Figure 2: DeepSort flow chart

The DeepSort algorithm is widely used in practical applications, the core is 2 algorithms: the Kalman filter and the hungarian algorithm. The implementation is relatively simple, often used in the combination for first detection, and then tracking.

Kalman filtering is the most efficient state-optimal estimation algorithm, which can predict the position about tracking target in consecutive frames based on the vector and matrix representing the target motion state. We usually use an 8-dimensional state vector to represent the motion state regarding a target, as shown in Eq. (1), containing the center coordinates, aspect ratio, height, and their respective velocities concerning a pair about targets.

$$\left[x, y, a, h, v_x, v_y, v_a, v_h\right] \tag{1}$$

Motion state estimation is achieved by using an 8-dimensional state space, where (x, y) is the center position concerning tracking target, a is the aspect ratio, h is the height, (v_x, v_y, v_a, v_h) is the velocity information about the target, respectively, and the remaining four variables represent the initial elements, (x, y, a, h) represents the original state for the target. DeepSort uses the Kalman filter algorithm to use the above 8-dimensional state space as the observation model for target state. The final prediction is (x, y, a, h), it is also the information about the position of the detection box.

The role of the Kalman filter in DeepSort is to perform the estimated prediction about foreground object state. The state refers to the parameters related to the box (the presence of target object to be detected in the box), including the location for the center, height, aspect ratio, etc. The Kalman filter can predict the updated trajectory. The Kalman filter can predict the updated trajectory. DeepSort also uses the martingale distance when updating the trajectory state, which is used as a correlation metric and is calculated as the Eq. (2):

$$d^{(1)}(i,j) = \left(d_j - y_i\right)^T S_i^{-1} \left(d_j - y_i\right) \tag{2}$$

where d_j represents the position of the detection frame of the jth target, y_i represents the position of the target predicted by the ith tracker, and S_i^{-1} represents the covariance between d_j and y_i.

The mahalanobis distance takes the form for calculating the standard deviation between the position concerning TD frame and the position predicted by the tracker, thereby measuring the uncertainty of the state estimation. The Hungarian algorithm serves to perform the association between the data, which is also a key step in matching. Here the Hungarian algorithm uses the cost matrix for matching between detections. The hungarian algorithm based on the detection frame position and IOU, makes DeepSort further enhanced in terms about efficiency compared to the SORT algorithm.

2.3 ECA-Net Modules

The ECA attention module is an ultra-lightweight attention module. Its module structure is shown in Fig. 3. where H, W, and C denote the height, width, and number of channels for the feature map, respectively. It avoids dimensionality reduction operations and uses a cross-channel approach to reconstruct channel feature map information and assign features. It uses dynamic convolution kernels for convolution and learns the importance of different channels. ECA attention module can improve the proportion and weight of small target feature maps, efficiently realize local cross-channel interaction, extract inter-channel dependencies, and enhance the feature extraction capability regarding network, which can maintain performance while significantly reducing model complexity. In this way, the network can better focus on vehicle target features during the detection process, and the attention on secondary information is reduced. The reverse direction about feature maps is suppressed. In addition, ECA-Net has fewer parameters, the computational complexity will not be increased, and the algorithm performance can be significantly improved.

After a channel-level global average pooling (GAP) without dimensionality reduction, a one-dimensional sparse convolution is used to obtain the interaction between current channel with its k neighboring channels. k is the adaptive convolution kernel size, which is determined by ψ(C), and its equation is shown in Eq. (3), where γ and b are linear fitting nonlinear parameters, $|t|_{odd}$ denotes the nearest odd number to t. σ is the sigmoid activation function, the sigmoid function generates the channel weights, combines the original input features with the channel weights, and then obtains the features with channel attention. Important channels will be amplified and unimportant channels will

be suppressed to ensure the integrity for extracted target features. This allows the detect and track models to more accurately locate feature areas. The use of ECA-Net Modules can make the proposed method focus more on the target features to be detected and recognized, ignoring the interference of invalid features.

$$k = \psi(C) = \left| \frac{\log_2^{(c)}}{\gamma} + \frac{b}{\gamma} \right|_{odd} \tag{3}$$

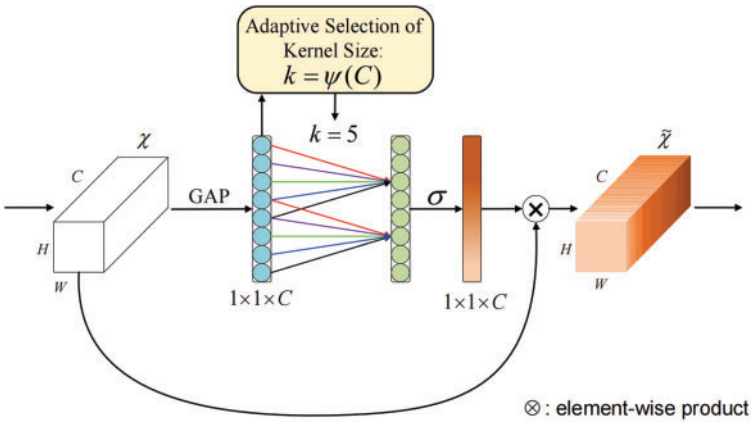

Figure 3: ECA-Net structure diagram

3 Implementation Detail

3.1 YOLOv5s Algorithm Improvement

Our improved YOLOv5s is the fastest training network of the four YOLOv5 versions. The main improvements made by this paper to the YOLOv5s network model are shown in Fig. 4. The first improvement is to add the ECA-Net attention mechanism to the C3 module of the network of the YOLOv5 backbone, which can be seen at the mark ① in the model diagram. Then, the ECA-Net attention mechanism is also added before the up-sampling of the neck feature pyramid, which can be seen at label ② in the model diagram. Finally, EIOU is used as the loss function of target frame regression. It can be seen that the improved YOLOv5 algorithm is called the YOLOv5-Block algorithm at label ③ in the model diagram.

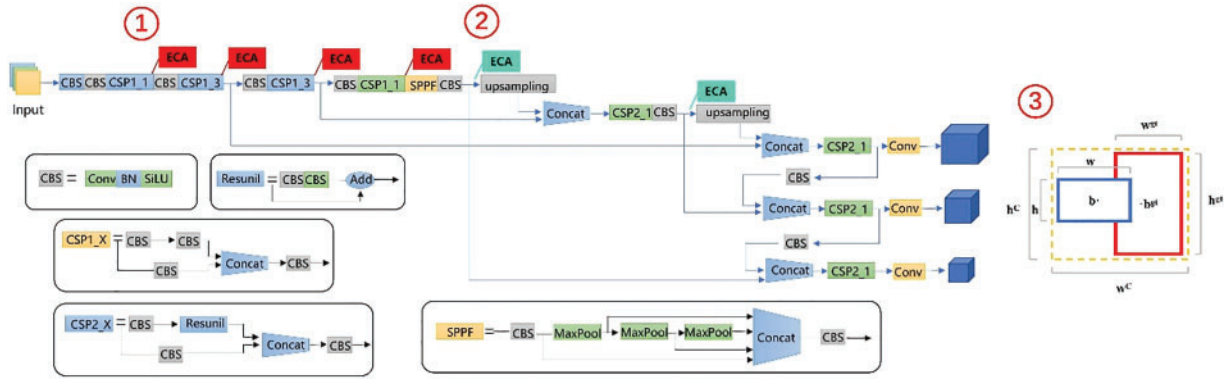

Figure 4: Schematic diagram of algorithm improvement

3.2 YOLOv5-Block Architecture

This paper improves the original feature fusion stage about YOLOv5s network. Since the color and shape regarding vehicle targets are similar, clear and distinctive features are difficult to extract, so this article adds ECA-Net behind the C3 module of the backbone network. The C3 module is mainly used to extract features from images. ECA-Net assigns different weights to different convolution channels to strengthen the feature extraction of the backbone network, so that the special vehicle target features are highlighted and the recognition accuracy is improved.

When the majority vehicle vehicles cause mutual occlusion, after the neck up-sampling process, the depth features regarding different targets will be merged, resulting in the model being unable to distinguish special vehicles. ECA-Net is added before the up-sampling concerning neck feature pyramid, which improves the utilization of feature information in the backbone network and the detection for special vehicles. By adding the ECA-Net-attention module, the effect of dimensionality reduction on the learning channel is avoided, the complexity about algorithm is greatly reduced, and the model is more comprehensive in extracting image features. The YOLOv5-Block network structure is shown in Fig. 5.

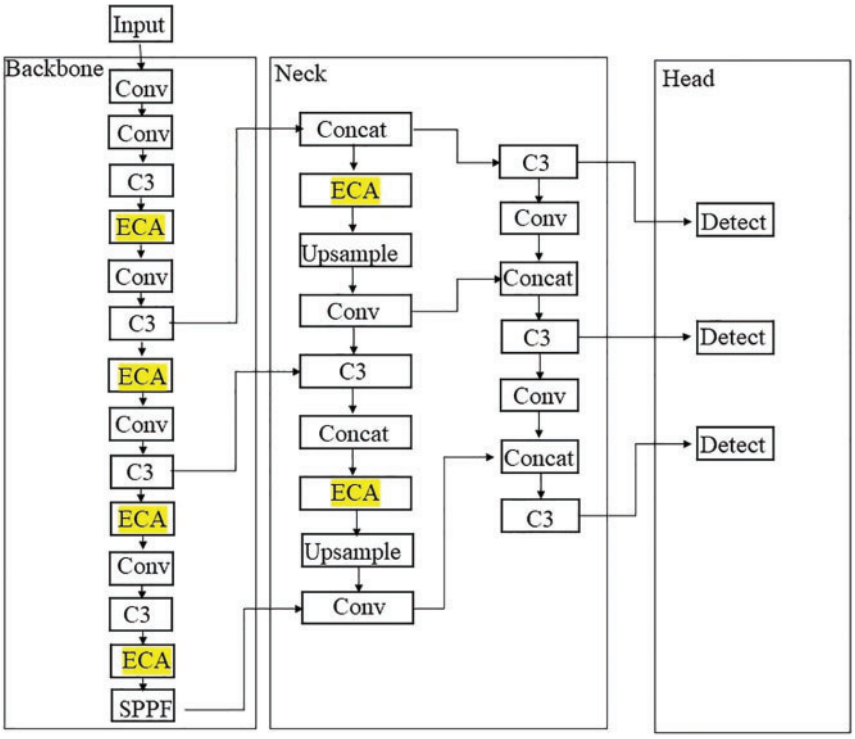

Figure 5: YOLOv5-Block structure diagram

3.3 Loss Function Improvement

The loss function evaluates the training results concerning model. The training results can be used to adjust the network weights and accelerate the convergence of it. The difference between the model and the actual data can be reflected. However, CIOU only takes the aspect ratio as an influencing factor. The aspect ratio is a relative value, and it does not consider the case that the target frame and the detection frame have the same aspect ratio but different values. The optimization regarding model is affected.

EIOU not only takes into account the center point distance and aspect ratio, but also takes into account the real differences in the width and height of the target frame, which makes the target frame regression process more stable and more accurate in convergence. The EIOU loss function is shown in Eq. (6), and The CIOU loss function is shown in Eqs. (4) and (5), where: b and b^{gt} denote the centroids for detection frame and the target frame, respectively; w and h denote the width and height about detection frame and the target frame, respectively; ρ denotes the Euclidean distance between the two centroids; c is the slope distance regarding smallest rectangle enclosing the detection frame and the target frame; α is a balance parameter that is not involved in the gradient calculation; v is the parameter used to measure the consistency for aspect ratio.

$$CIoU = IoU - \rho^2\left(b, b^{gt}t\right)\big)/c^2 - \alpha v \tag{4}$$

$$\alpha = \frac{v}{(1 - IoU) + v} \tag{5}$$

$$L_{EIoU} = 1 - IoU + \left(\rho^2\left(b, b^{gt}\right)\right)/c^2 + \left(\rho^2\left(w, w^{gt}\right)\right)/\left(C_w^2\right) + \left(\rho^2\left(h, h^{gt}\right)\right)/\left(C_h^2\right) \tag{6}$$

3.4 YOLOv5-Block Combined with DeepSort

In this paper, YOLOv5-Block is used to train the vehicle detection model. The multi-target detection results from the video are used as real-time input to the DeepSort tracker. The accuracy of the detection results is high, which makes up for the shortcomings about DeepSort's algorithm. Thus, the multi-targets in the video are tracked completely and accurately in real-time. The specific flow is shown in Fig. 6.

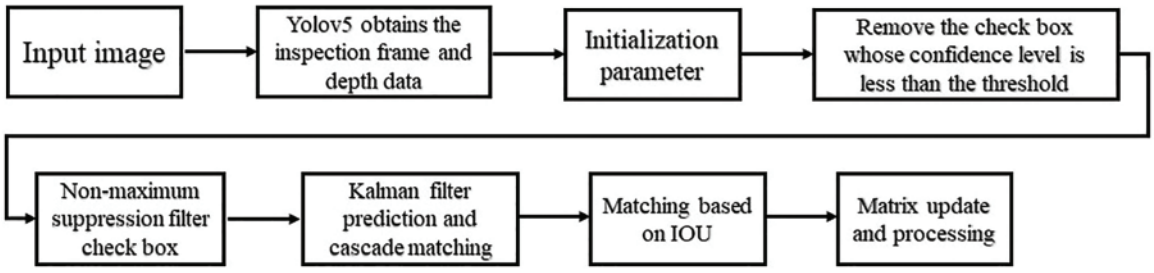

Figure 6: YOLOv5-Block combined with DeepSort algorithm flow chart

First, each frame regarding input video frame is processed, and the YOLOv5-Block TD algorithm reads the position about detection frame for vehicle to be detected in the current video frame and the depth characteristics concerning image blocks in each detection frame and filters the detection frames according to the confidence level, i.e., the detection frames with insufficient confidence are removed. Then the Kalman filter is used to predict the position of the vehicle in the current frame, and the accuracy is improved by cascade matching. Then followed by matching between their tracking and detection frames based on IOU for trackers on non-cascading matching and tracking on unconfirmed state with detection on non-cascading matching. The last step is to do the parameter update for the trackers on cascade matching.

4 Experiment

4.1 Experimental Environment and Parameter Settings

The experiment is conducted under a 64-bit Windows 10 operating system with central processing unit (CPU) E5-2670 v3 @ 2.30 GHz running on 16 GB of random-access memory and Nvidia-1060

6 GB graphics processing unit (GPU), using the GPU version of PyTorch, choosing CUDA version 11.3, Python 3.9 programming language, and pycharm are used for the experiments. The experiments are performed with adaptive anchor, mosaic data enhancement, batch size set to 4, epoch set to 300, images normalized to 640 × 640 resolution, learning rate of 0.01, and optimizer using stochastic gradient descent.

4.2 Dataset Description

Due to the lack of publicly available image datasets for special vehicles (such as tanks and other armored vehicles), and considering the confidentiality about military equipment and entering the battlefield environment to collect image data for special vehicles (such as tanks and other armored vehicles), there is a serious risk of human life and security risks. We decided to independently produce a dataset with 3886 pictures including multiple armored vehicle targets through network resources. The dataset includes all kinds of complex scenes (such as extreme conditions such as mountains, rivers, deserts, and heavy snow) and multi-scale and various forms of special vehicle pictures. The display concerning special vehicle datasets in virtual scenes is shown in Fig. 7.

Figure 7: Special vehicle dataset

All images were saved in JPEG format with a high resolution of 1920 × 1080, divided into a training set and test set according to the ratio of 8:2, and the Labelimg annotation tool was used to mark the real target frame for each image. Mosaic data enhancement, initialization, normalization, and other processing were performed on all data before training. To measure the improved model's recognition regarding vehicle targets under occlusions, we especially collected special vehicle images under occlusions, including different angles, different scenes, different vehicle numbers, and other conditions. A typical vehicle occlusion image is shown in Fig. 8.

Figure 8: Typical occlusion image

4.3 Evaluation Indicators

To accurately evaluate the testing effect of the experiment, the precision rate (P) and AP, and correct number (N) are used as the evaluation criteria about TD model, which is calculated by Eq. (9), where: true positives (TP) sample predicted by the sample; false positives (FP) sample predicted by the sample; false negatives (FN) sample predicted by the sample. N stands for the number of images correctly recognized under occlusion.

$$p = \frac{TP}{FP + TP} \tag{7}$$

$$R = \frac{TP}{FN + TP} \tag{8}$$

$$AP = \int_0^1 P(R)\, dR \tag{9}$$

Evaluate the multi-target tracking effect by using the metrics defined by multiple objects tracking (MOT) Challenge: In this paper, three metrics are selected as multi-target tracking effect evaluation metrics: MOTA, MOTP, and IDSW. IDSW is the number of target ID transformations, the larger the value, the more frequent the target ID transformation, and the weaker the tracking and re-identification ability of the tracker, MOTA reflects the accuracy for multi-target tracking, and MOTP reflects the accuracy regarding tracked target position.

$$MOTA = 1 - \frac{\sum_t (m_t + n_t + s_t)}{\sum_t g_i} \tag{10}$$

$$MOTP = \frac{\sum_{i,t} d_i^t}{\sum_t c_t} \tag{11}$$

In Eq. (10), t denotes the tth frame; m_t denotes the number of missed targets in the tth frame; n_t denotes the number regarding false targets in the tth frame; s_t denotes the number for identity switching in the tth frame; g_t denotes the total number of targets appearing in the tth frame; d_i^t denotes the calculated distance between the predicted position and the real position of target i in the tth frame; c_t denotes the number about targets successfully matched between the predicted and labeled trajectories in the tth frame.

MOTA is not related to the estimation accuracy of object position, but MOTA is a relatively accurate measure of the TA's capability in detecting objects and tracking trajectories.

4.4 Experimental Results and Analysis

To verify the effectiveness for improved YOLOv5 algorithm proposed in this paper. Experiments were conducted on a homemade vehicle dataset. The experimental results are presented in the following Table 1. As can be seen from Table 1, the recognition accuracy of the YOLOv5 model has been very high after adding the attention mechanism. YOLOv5 shows a slight decrease in recognition accuracy with the addition of the CBAM attention mechanism, but the number of recognized images increases by 21 for the occlusion case. After adding the ECA attention mechanism to YOLOv5, there is no change in the recognition accuracy, but for the occlusion case the number of recognized pictures increases by 26 pictures, therefore, the addition of the ECA attention mechanism to the YOLOv5 structure is useful for improving the occlusion case that arises when the vehicle targets are close to each other. Next, this paper investigates the effect about type regarding loss function on target recognition. Based on YOLOv5 algorithm which has added ECA attention mechanism, three types of loss functions, wise IOU (WIOU), soft IOU (SIOU) and EIOU, are added, respectively, and it can be seen that there is a great improvement in the recognition concerning occlusion situation in all three cases. Among them, the model with the addition for EIOU's loss function has the greatest improvement in the number of correctly recognized pictures for the occlusion situation, which reaches 57 pictures. Although the precision rate of the YOLOv5-Block has decreased, the number of correctly identified vehicle target images has increased significantly. But it is acceptable. For the specific military application scenario in this paper, we are more concerned with the number of images that can correctly identify the vehicle target rather than the precision rate. This means that we can find more enemy target threats on the battlefield. The test results concerning typical occlusion images are shown in

Fig. 9. The algorithm YOLOv5-Block has more information fusion channels by adding ECA-Net to the backbone and neck networks. The algorithm can identify and locate vehicles more accurately, and dense and obscured targets can be identified more accurately. Therefore, the experiment verifies the effectiveness for the improved model in this paper.

Table 1: Improved YOLOv5 experiments

Models	P	AP	N
YOLOv5	0.875	0.995	23
YOLOv5 + CBAM	0.885	0.991	44
YOLOv5 + ECA	0.555	0.995	49
YOLOv5 + ECA + WIOU	0.658	0.995	52
YOLOv5 + ECA + SIOU	0.708	0.995	41
YOLOv5-Block	0.708	0.995	**57**

Figure 9: Test images

The following verifies the tracking effectiveness of the improved model. The experiments are conducted following the criteria concerning MOT challenge for video sequence testing. In this paper, the original YOLOv5 and YOLOv5-Block are combined with the DeepSort algorithm, respectively. Both algorithms were tested in a set of test videos. The advantages and disadvantages of the improved YOLOv5 algorithm are analyzed. The experimental results and tracking effects are shown in Table 2, Figs. 10, and 11. As can be seen from Table 2, in the tests of different virtual simulation video sequences, compared with YOLOv5+DeepSort algorithm, MOTA and MOTP of YOLOv5-Block+DeepSort algorithm are greatly improved, and IDSW decreases significantly. This shows that the algorithm proposed in this paper is effective for trajectory tracking in the case of vehicle occlusion.

Table 2: Multi-target tracking evaluation experiments

Algorithm type	Video sequence	MOTA	MOTP	IDSW
YOLOv5 + DeepSort algorithm	01	56.51%	21.14%	3
	02	95.39%	15.24%	0
	Average	75.95%	18.19%	1.5
YOLOv5-Block + DeepSort algorithm	01	74.71%	23.57%	2
	02	98.58%	16.31%	0
	Average	**86.65%**	**19.94%**	**1**

(a) (b)

Figure 10: The tracking process of video sequence 01 frames 50, 108, 117, 139 (a) tracking process before the improvement (b) tracking process after the improvement

Fig. 10a is the video frame of the tracking process before the improvement for video sequence 01, and Fig. 10b is the video frame for tracking process after the improvement, in which the vehicle marked with the red dot is the main tracking target. It can be seen in (a) that the YOLOv5+DeepSort algorithm can continue to track the two vehicles before they meet. In frame 117, when two vehicles

meet, the original algorithm cannot identify targets that block each other, which leads to errors in the TA, and their target boxes are merged into target boxes with ID 2. In the video frame concerning the improved algorithm, after the two vehicles meet in frame 117, the TA will not be able to recognize each other. The algorithm proposed in this paper can still identify mutually occluded targets and keep the ID unchanged in the tracking process. The comparison of video frames (a) and (b) demonstrates the efficacy of the YOLOv5-Block+DeepSort algorithm in accurately identifying and tracking mutually occluded vehicle targets.

(a) (b)

Figure 11: Tracking process of video sequence 02 frames 37, 67, 153, 160 (a) tracking process before the improvement (b) tracking process after the improvement

Fig. 11a is the video frame of the tracking process for video sequence 02 before the improvement; Fig. 11b is the video frame about the tracking process after the improvement. Frames 37 and 67 can be seen in (a). The algorithm before the improvement can only identify the most complete tracking vehicle, while the target is obscured by the vehicle and another distant vehicle cannot be identified. In Frames 153 and 160 of the improved algorithm (b), the obscured target and the distant target were identified and continuously and accurately tracked. In Frames 153 and 160, the original algorithm could not accurately track the vehicle target under crowded occlusion, resulting in merging and position confusion concerning target frame. In the improved algorithm, the YOLOv5-Block+Deepsort algorithm can carry out an accurate and continuous tracking process, regardless of whether the target is under congested occlusion or a distant occluded target. The reason is speculated that the YOLOv5-Block TD algorithm adopts the ECA-Net structure in both the backbone network and the neck network. ECA-Net enhances the boundary features of dense vehicles through fused channel and weighted channel features. Feature information regarding occluded targets is highlighted. The detection algorithm can correctly identify densely packed vehicles. Due to the accuracy for the detection algorithm's positioning, the Deepsort algorithm continuously updates the tracking position frame developed on the accurate detection frame during the tracking process, maintaining the accuracy

and stability about tracking process. Therefore, the YOLOv5-Block+Deepsort algorithm proposed in this article has significantly improved the tracking process at multiple scales including large vehicle target occlusion.

4.5 Comparative Experiments

To further verify the performance for our proposed method, we conducted a comparison experiment with the latest mainstream vehicle tracking and detection model YOLOv10. The experimental results of comparison between the proposed method and YOLOv10 are shown in Table 3.

Table 3: Comparative experimental results

Models	P	AP	N
YOLOv10	0.882	0.993	55
Ours	0.708	0.995	57

As can be seen from Table 3, the P concerning YOLOv10 is higher than that of our proposed method, but our method is slightly better than YOLOv10 in terms of AP and the number of images correctly recognized under occlusion N.

The visualized comparison results are shown in Fig. 12 below. As can be seen in Fig. 12, our method accurately and completely identifies all specific vehicle targets.

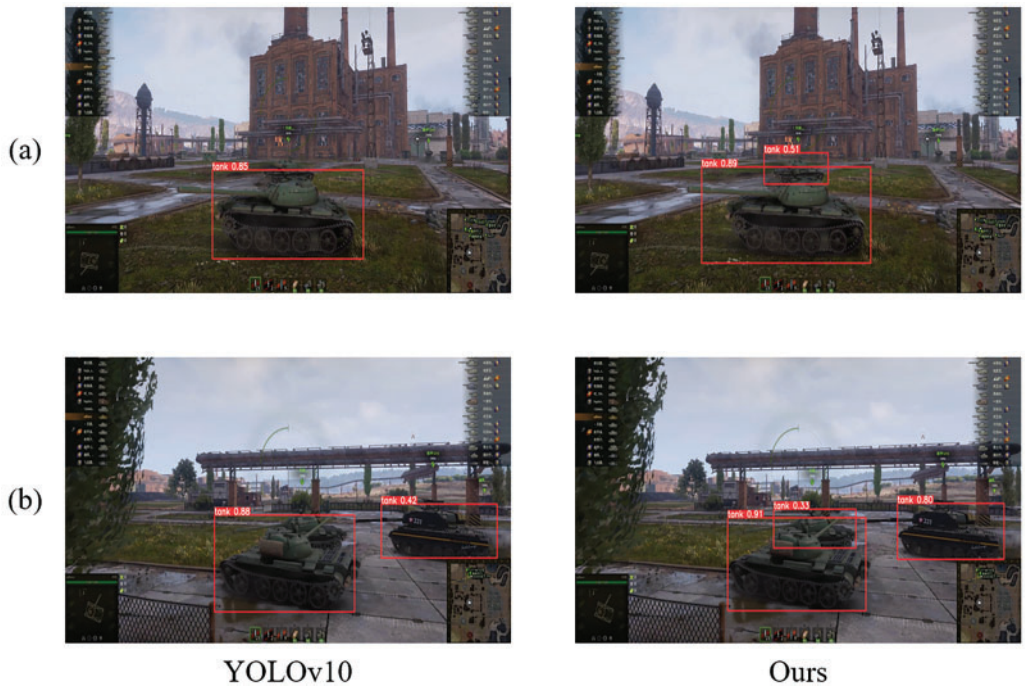

Figure 12: Comparison of visual results (a) Scenario #1 (b) Scenario #2

4.6 Experimental Comparison of Application Scenarios

We selected a video of multiple special vehicles driving in a virtual real environment and occluded each other during the driving process. We used our proposed method YOLOv5-Block+DeepSort algorithm and the mainstream tracking algorithm SFFSORT to conduct experiments on the video sequence, respectively.

The experimental comparison results concerning frame 10 and frame 23 about the video sequence are captured and shown in Fig. 13. From Fig. 13a,b, it can be seen that at the beginning about video, the SFFSORT algorithm cannot correctly identify the two special vehicles that are occluded due to their adjacent, and the SFFSORT algorithm identifies the two vehicles for a certain type into one. At this time, as the video plays, the SFFSORT algorithm can only draw a trajectory curve. However, our proposed method correctly identified the target boxes of two special vehicles and drew two correct trajectory curves.

(a) SFFSORT (b) proposed method

Figure 13: Comparison of experimental results of SFFSORT algorithm and proposed method

5 Conclusion

In this paper, a YOLOv5-Block+Deepsort algorithm is proposed to solve the problem of failing to accurately and continuously track a target in a virtual simulation scenario due to the mutual occlusion for different vehicles during traveling. By incorporating the ECA-Net attention mechanism into the YOLOv5 network structure and improving the loss function, the YOLOv5-Block algorithm improves the detection correctness of typical occlusion datasets by a factor concerning 1.48 over no improvement. It can effectively identify obscured vehicle targets. After the YOLOv5-Block algorithm was fused with Deepsort TA, MOTA and MOTP improved by 10.7 and 1.75 percentage points in the video tracking test, while IDSW declined. This shows that the tracking accuracy of the YOLOv5-Block+Deepsort algorithm is improved and the identity-switching problem caused by occlusion is reduced. Large vehicles can be tracked consistently and accurately under multiple scales of occlusion. Therefore, the YOLOv5-Block+DeepSort algorithm has higher accuracy and better continuous tracking effect in vehicle tracking in both simple and complex scenes.

Although the method proposed in this paper has improved the detection accuracy and tracking effect of special vehicles to a certain extent, there is still room for further improvement in the parameters and lightweight of the network. Whether it is suitable for the traffic scenario of civilian vehicle type needs to be studied in future work.

Acknowledgement: The authors would like to express their heartfelt gratitude to the editors and reviewers for their detailed review and insightful advice.

Funding Statement: This work was supported in part by the National Key R&D Program of China under Grant 2022YFB2703304; in part by the National Natural Science Foundation of China under Grant 52074064; in part by the Fundamental Research Funds for the Central Universities under Grant N2404013, N2404015.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Le Wang, Zhenni Li; data collection: Hanquan Zhang, Zhengmin Gu; analysis and interpretation of results: Mingyuan Zhai; draft manuscript preparation: Dong Xiao. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data and codes that support the findings of this study are available from the corresponding authors upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] L. Jiao, D. Wang, Y. Bai, P. Chen, and F. Liu, "Deep learning in visual tracking: A review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 9, pp. 5497–5516, Sept. 2023. doi: 10.1109/TNNLS.2021.3136907.

[2] L. Kalake, W. Wan, and L. Hou, "Analysis based on recent deep learning approaches applied in real-time multi-object tracking: A review," *IEEE Access*, vol. 9, pp. 32650–32671, 2021. doi: 10.1109/ACCESS.2021.3060821.

[3] M. Lao, X. Chen, F. Lin, G. Qin, W. Liu and Y. Zhou, "Visual target detection and tracking framework using deep convolutional neural networks for micro aerial vehicles," in *IEEE 14th Int. Conf. Cont. Automat. (ICCA)*, Anchorage, AK, USA, 2018, pp. 276–281.

[4] S. Yan, Y. Fu, W. Zhang, W. Yang, R. Yu and F. Zhang, "Multi-Target instance segmentation and tracking using YOLOV8 and BoT-SORT for video SAR," in *5th Int. Conf. Elect. Eng. Inform. (EEI)*, Wuhan, China, 2023, pp. 506–510.

[5] L. Li and Y. Liang, "Deep learning target vehicle detection method based on YOLOv3-tiny," in *IEEE 4th Adv. Inform. Manag., Commun., Elect. Automat. Cont. Conf. (IMCEC)*, Chongqing, China, 2021, pp. 1575–1579.

[6] H. Yu, K. Meier, M. Argyle, and R. W. Beard, "Cooperative path planning for target tracking in urban environments using unmanned air and ground vehicles," *IEEE/ASME Trans. Mechatron.*, vol. 20, no. 2, pp. 541–552, Apr. 2015. doi: 10.1109/TMECH.2014.2301459.

[7] X. Cheng, J. Zhou, P. Liu, X. Zhao, and H. Wang, "3D vehicle object TA based on bounding box similarity measurement," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 12, pp. 15844–15854, Dec. 2023. doi: 10.1109/TITS.2023.3278378.

[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, 2014, pp. 580–587.
[9] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, 1 Sep. 2015. doi: 10.1109/TPAMI.2015.2389824.
[10] G. Gkioxari, R. Girshick, and J. Malik, "Contextual action recognition with R*CNN," in *IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, 2015, pp. 1080–1088.
[11] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 1 Jun. 2017. doi: 10.1109/TPAMI.2016.2577031.
[12] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, real-time object detection," in *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 779–788.
[13] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Eur. Conf. Comput. Vis. (ECCV)*, Cham, Springer, 2016, pp. 21–37.
[14] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, 2017, pp. 6517–6525.
[15] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.
[16] A. Bochkovskiy, C. Y. Wang, and H. Y. Mark Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
[17] Y. Ma, "PANet: Parallel attention network for remote sensing image semantic segmentation," in *ISCTT 6th Int. Conf. Inform. Sci. Comp. Technolo. Transportati.*, Xishuangbanna, China, 2021, pp. 1–4.
[18] L. Meng and X. Yang, "A survey of object tracking algorithms," (in Chinese), *Acta Automat. Sinica*, vol. 45, no. 7, pp. 1244–1260, 2019.
[19] W. Luo, J. Xing, A. Milan, X. Zhang, W. Liu and T. K. Kim, "Multiple object tracking: A literature review," 2022, *arXiv:1409.7618*.
[20] X. Li *et al.*, "A comprehensive review of deep learning-based object tracking algorithms," (in Chinese), *J. Image Graph.*, vol. 24, no. 12, pp. 2057–2080, 2019.
[21] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1409–1422, Jul. 2012. doi: 10.1109/TPAMI.2011.239.
[22] N. Ali and G. Hassan, "Kalman filter tracking," *Int. J. Comput. Appl.*, vol. 89, no. 9, pp. 15–18, 2014.
[23] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-Speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, 1 Mar. 2015. doi: 10.1109/TPAMI.2014.2345390.
[24] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *IEEE Int. Conf. Image Process. (ICIP)*, Phoenix, AZ, USA, 2016, pp. 3464–3468.
[25] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *IEEE Int. Conf. Image Process. (ICIP)*, Beijing, China, 2017, pp. 3645–3649.
[26] Z. L. Tian, M. N. A. Wahab, M. F. Akbar, A. S. A. Mohamed, M. H. M. Noor and B. A. Rosdi, "SFFSORT multi-object tracking by shallow feature fusion for vehicle counting," *IEEE Access*, vol. 11, pp. 76827–76841, 2023. doi: 10.1109/ACCESS.2023.3297190.
[27] J. Li, H. Song, Z. Zhang, J. Hou, and F. Wu, "Multi-object vehicle tracking and trajectory optimization based on video," (in Chinese), *Comput. Eng. Appl.*, vol. 56, no. 5, pp. 194–199, 2020.
[28] L. Liu, S. Zhao, and W. Guo, "A method for traffic flow statistics based on YOLO recognition and mean shift tracking," (in Chinese), *Manuf. Automat.*, vol. 42, no. 2, pp. 16–20, 2020.
[29] L. Jin, Q. Hua, B. Guo, X. Xie, F. Yan and B. Wu, "Front vehicle multi-object tracking based on optimized DeepSort," (in Chinese), *J. Zhejiang Univ. (Engineer. Sci. Ed.)*, vol. 55, no. 6, pp. 1056–1064, 2021.

[30] M. A. Bin Zuraimi and F. H. Kamaru Zaman, "Vehicle detection and tracking using YOLO and DeepSORT," in *IEEE 11th IEEE Symp. Comput. Appl. & Ind. Elect. (ISCAIE)*, Penang, Malaysia, 2021, pp. 23–29.

[31] Z. Zhao, Z. Ji, Y. Yao, Z. He, and C. Du, "Enhanced detection model and joint scoring strategy for multi-vehicle tracking," *IEEE Access*, vol. 11, pp. 30807–30818, 2023. doi: 10.1109/ACCESS.2023.3262466.