Journal of Computers Vol. 34 No. 1, February 2023, pp. 175-186
doi: 10.53106/199115992023023401013

Violation Behavior Detection for Non-motor Vehicles

Wei-Yi Jing[1], Zhong-Jie Zhu[1*], Yong-Qiang Bai[1], Long Li[1], Wei-Feng Cui[2], Wen-Bo Yu[1]

[1] Ningbo Key Lab of DSP, Zhejiang Wanli University, Ningbo 315000, China
weiyijing_zwu@163.com, zhongjiezhu@yeah.net, byq-163@163.com, 740605729@qq.com, 3473388492@qq.com
[2] Physical Engineering College, Zhengzhou University, Zhengzhou 450000, China
940104749@qq.com

Received 25 April 2022; Revised 28 July 2022; Accepted 10 August 2022

Abstract. Non-motor vehicles are widely used in the urban and rural transportation system for their portability, but the related violations also occur frequently and are difficult to be supervised intelligently, considering their colossal quantity, various styles, and small volumes. To solve this problem, this paper presents a non-motor vehicle violation detection algorithm with efficient target detection and deliberate logical calculation. A target detection network with high speed and accuracy is constructed firstly by fusing two different types of attention mechanism. Specifically, the Squeeze-and-Excitation Network is employed to optimize the extraction of local features, which can effectively reduce the error rate of target detection. Meanwhile, the Transformer Network is adopted to strengthen the extraction of global features, which can improve the target location performance for target tracking in high-density scenarios. As the global and local features are integrated with attention mechanism, the proposed network can accurately identify and locate the numerous small targets in real-time, and avoid identity switching caused by target occlusion. Finally, the violation of non-motor vehicles is recognized in real-time by constructing the logical calculation between the target features and their motion trajectories. Experiments on datasets show that the detection accuracy of the proposed algorithm is better than the current mainstream algorithms, especially the accuracy of small targets such as Head and Helmet is higher than 92.2%. And our ID Switch is dropped by more than 60% compared with the classical Deep SORT algorithm. In real-life scenarios, the proposed algorithm also shows excellent accuracy and real-time performance for non-motor vehicle violations.

Keywords: target detection, tracking, attention mechanism, non-motor vehicles, violations

1 Introduction

Non-motor vehicles (bicycles, tricycles, etc.) are one of the typical traffic modes for short-distance travel in urban and rural areas, considering their low carbon, convenience, and other characteristics. In particular, they have also been widely used in the “last mile” logistics distribution. However, the mixed traffic flow with electric bicycles and other non-motor vehicles performances more complex than the motorized traffic flow because of its colossal quantity, various styles, and small volumes. As a result, non-motor vehicle violations still rely mainly on the human investigation by traffic police, which is time-consuming, labor-consuming, and ineffective. Hence, it is imperative to apply an efficient and automatic non-motor vehicle violations detection algorithm for the construction and development of intelligent transportation.

Up to now, there is little related literature and information on non-motor vehicle violation detection. Strictly speaking, it belongs to the field of target detection and tracking for small objects, combined with subsequent logical judgment. For target detection, the mainstream algorithms are mainly based on deep learning. They can be roughly divided into two categories according to the different network structures, which are candidate region-based algorithms and regression-based algorithms [1-3]. These algorithms have good detection accuracy, but poor speed as each candidate region needs to perform convolution operations [4]. The regression-based algorithms directly carry out position regression on the extracted features to replace the candidate region, which can significantly improve the detection speed and is represented by Single Shot MultiBox Detector (SSD) series [5-6] and You Only Look Once (YOLO) series algorithms [7-10]. Among them, the YOLOv5 network uses the CSP1_X structure to alleviate inference calculation and achieves a good balance between accuracy and speed. In addition, Feature Pyramid Networks (FPN) [11] and Path Aggregation Network (PAN) [12] are further used

[*] Corresponding Author

to transmit deep semantic features, and shallow position information to improve detection accuracy. These target detection algorithms are still insufficient for feature extraction of small targets such as Helmets and Heads. And these small targets are realized for subsequent target tracking, which directly affects the accuracy and real-time violation detection. For target tracking, the mainstream algorithm is back-end tracking optimization algorithms based on the Hungarian, including Simple Online and real-time Tracking (SORT) [13] and Simple Online and real-time Tracking with A Deep Association Metric (Deep SORT). Specifically, SORT focuses on frame-to-frame prediction and association, realizing efficient tracking based on ordinary CNN. Based on SORT, Deep SORT adds appearance information to achieve long-term occlusion target tracking.

Target detection shows great potential in detection accuracy and detection speed, and target tracking can better overcome the occlusion problem. Inspired by this, a non-motor vehicle violation detection algorithm with efficient target detection and logical calculation is proposed, which realizes the complementary advantages of target detection and target tracking based on deep learning to solve the problem of non-motor vehicle violation detection. The experimental results demonstrate the effectiveness of the proposed algorithm. The main contributions of this paper are summarized as follows:

1. Considering the high complexity of real traffic scenes and the performance of the detection model is very dependent on the database used for training, a new non-motor vehicle violation detection database is established to better match the proposed algorithm to achieve high performance.
2. A detection network combined with Transformer is constructed, which improves the connection of contextual features and performance of global features extraction. SE attention mechanism is integrated into the network to optimize the extraction of local features. The effective combination of Transformer and SE attention mechanism realizes the fusion of global features and local features, which improves the performance of the detection network for small size targets.
3. To solve the problem of accurately detecting small targets in the presence of occlusion, features of the motion trajectory of target are considered. Through the detection combined with the tracking to jointly realize non-motor vehicle violation detection.

The rest of this paper is organized as follows. Section 2 describes the related work. In Section 3, the proposed algorithm is introduced in detail. Then, the relevant experimental results and a comparative analysis are presented in Section 4. Finally, Section 5 concludes this paper.

2 Related Work

Non-motor vehicle violation detection is an important part of modern intelligent transportation. At present, some work has been carried out on the research, and which can be mainly divided into two categories: traditional detection algorithms and deep-learning based detection algorithms.

Traditional detection algorithms mainly rely on manual features to detect and track non-motor vehicle violations. Feng et al. combined the image difference method and fractional differentiation method to extract the contour of dynamic non-motor vehicles, and used skin color filter detection and geometric feature discrimination to filter the contour of non-motor vehicles to assist image processing. The adaptive mean shift method is proposed to detect the trajectory, so as to judge whether the current non motor vehicle is the same as the detected illegal non motor vehicle [14]. Silva et al. proposed using the background subtraction method to extract moving vehicle objects, combined with the use of local binary pattern to classify the extracted features. After classifying motorcycle objects, cut out the helmet part from 1/5 of the image, and classify them using HOG, Hough transform and LBP descriptor [15]. Wen et al. proposed a circular arc detection method based on improved Hough transform to detect helmets in a ATMs [16]. However, in the face of the increasingly complex traffic environment, the traditional detection algorithms have gradually been unable to meet the requirements in terms of detection accuracy. Complex detection scenes and occlusion will affect the reliability of traditional algorithms.

Deep-learning based detection algorithms have shown great potential in overcoming the above problems. Mistry et al. proposed a detection algorithm using YOLOv2 to detect motorcyclists wearing helmets and not wearing helmets. In order to achieve this purpose, the detection objects are divided into personnel detection, helmet detection and license plate detection. Use YOLOv2 to detect personnel and helmet objects, and use OpenALPR algorithm to detect license plate coordinates. The judgment process is that if no license plate is detected, it means that YOLOv2 detects not people on motorcycles, but pedestrians [17]. YOLOv2 maintains the speed advantage by introducing batch normalization, anchor box and high-resolution classifier. However, there are some problems such as low accuracy and inaccurate positioning. Hirota proposed a classification of helmet and non-helmet motorcyclists based on CNN, but helmets of different colors hinder the accuracy of detection [18].

However, while deep-learning based detection algorithms have shown some promising performance, there is still room for improvement. At present, there is no large-scale public database for non-motor vehicle violation detection to support further research. Algorithms relying solely on deep-learning based detection cannot solve the problem of accurate detection of occluded targets and continuous accurate tracking.

3 Proposed Algorithm

In this paper, a new database is established and a real-time detection algorithm is proposed for of non-motor vehicle violation detection, as is shown in Fig. 1. The proposed algorithm mainly includes three parts: target detection, target tracking, and judgment of a breach. In terms of target detection, multi-scale information perception based on Transformer is considered, while SE attention mechanism is combined to improve the performance of feature extraction for small targets. Based on target detection, the feature of target motion trajectory is added to jointly realize non-motor vehicle violation detection.

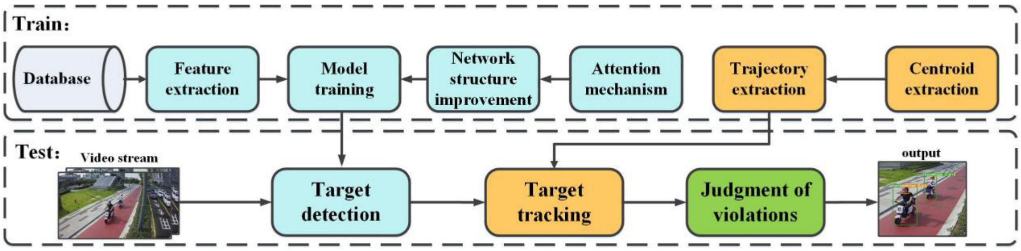


Fig. 1. The flow of the proposed algorithm

3.1 Establishment of The Database

To fully cope with the complex detection environment of traffic and to better fit the proposed algorithm, a new database for non-motor violation detection is established, partial pictures of the database is shown in Fig. 2. Considering that different lighting conditions, different backgrounds, and different scenarios may affect the accuracy of detection, the establishment of the database is mainly composed of four aspects, that is, pictures in low light and backlight environments, pictures in strong light environments, pictures with simple background, and pictures with complex background.

These pictures are collected by different traffic cameras, ensuring the authenticity and complexity of the traffic environment. The non-motor vehicles in the database are mainly two-wheeled vehicles and three-wheeled vehicles, which are the most typical non-motor vehicles in current urban traffic. In addition, these pictures include three types of violations of non-motor vehicle drivers not wearing helmets, illegally carrying people and driving in reverse. And the database has a total of 20,000 pictures, each with a size of 1920×1080.

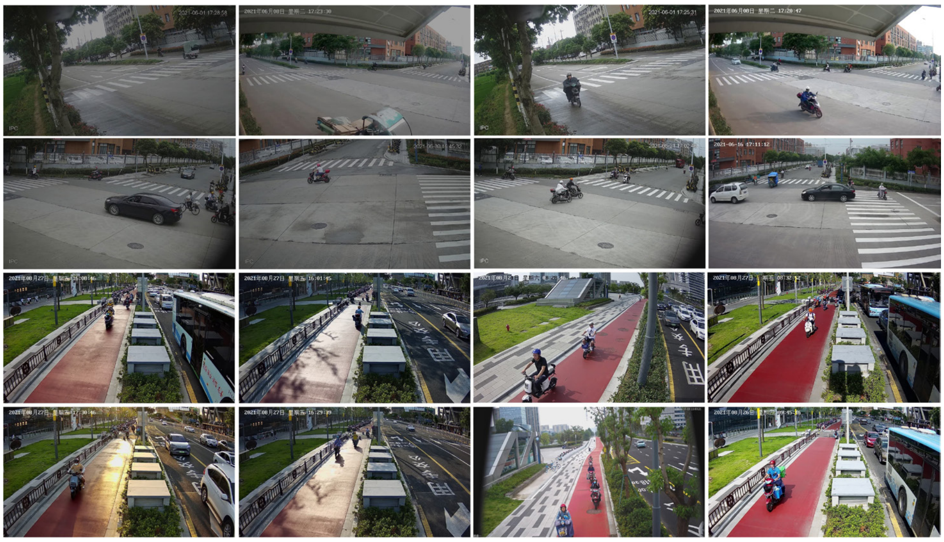


Fig. 2. Partial pictures of the database

To better serve the subsequent training, the targets in the dataset are labeled by LabelImg, as shown in Fig. 3. The labeled targets are divided into four categories, including head, helmet, two-wheeled vehicle, and three-wheeled vehicle. For the possible training loss due to ground truth errors in subsequent training, each target is labeled compactly. Among the database, the head and helmet are especially carefully labeled, which belong to small targets, so that the edge of the ground truth is close to the target contour. For the labeling of non-motor vehicles, the strategy of labeling the driver and non-motor vehicles together is adopted, which effectively enables the proposed algorithm to distinguish non-motor vehicles in motion from non-motor vehicles parked on the roadside. Besides, there are approximately 80,000 labels in total, and the ratio of the number of labels for the four categories is approximately 1:1:1:1.

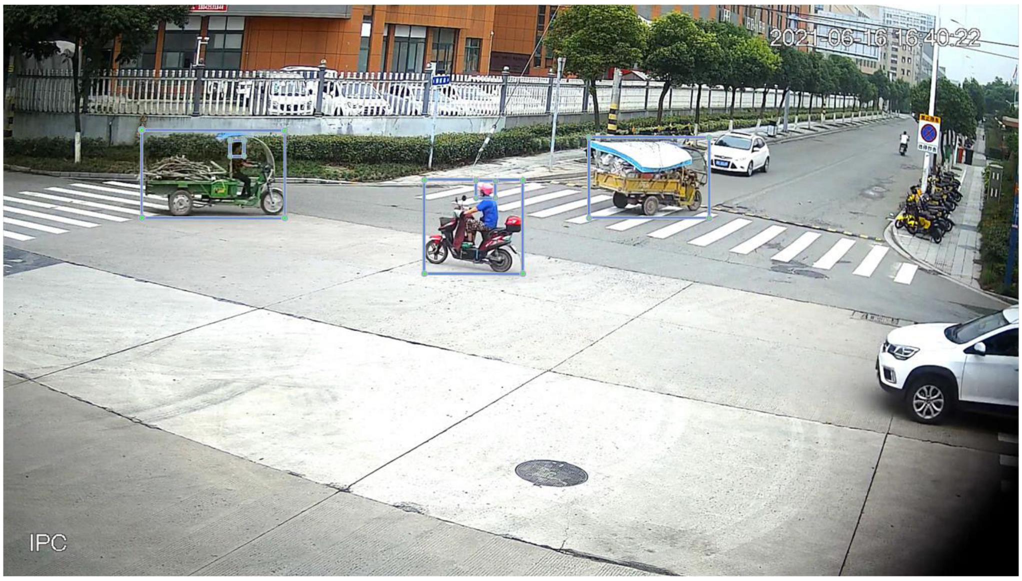


Fig. 3. Labels of the database

3.2 Target Detection

According to the labels of the database, four categories of targets to be detected are determined: Head, Helmet, Two-wheeled vehicle, and Three-wheeled vehicle. Then a detection network combined with a Transformer is constructed, linking the different scale receptive fields of the spatial pyramid features. In contrast, to make the detection network more integrated, the low-level and high-level features are merged. In addition, the SE attention mechanism is used to weight the output of the contextual features by the Transformer, which has the advantage that the importance of each feature channel can be obtained through automatic learning based on fewer calculations. In constructing the detection network, using the C3 structure based on the CONV module to enhance the performance of feature extraction without increasing the amount of calculation as much as possible. Finally, four different sizes of receptive fields are output to deal with the prediction of three different sizes of targets: large, medium, small and tiny. The block diagram of the proposed target detection network is shown in Fig. 4.

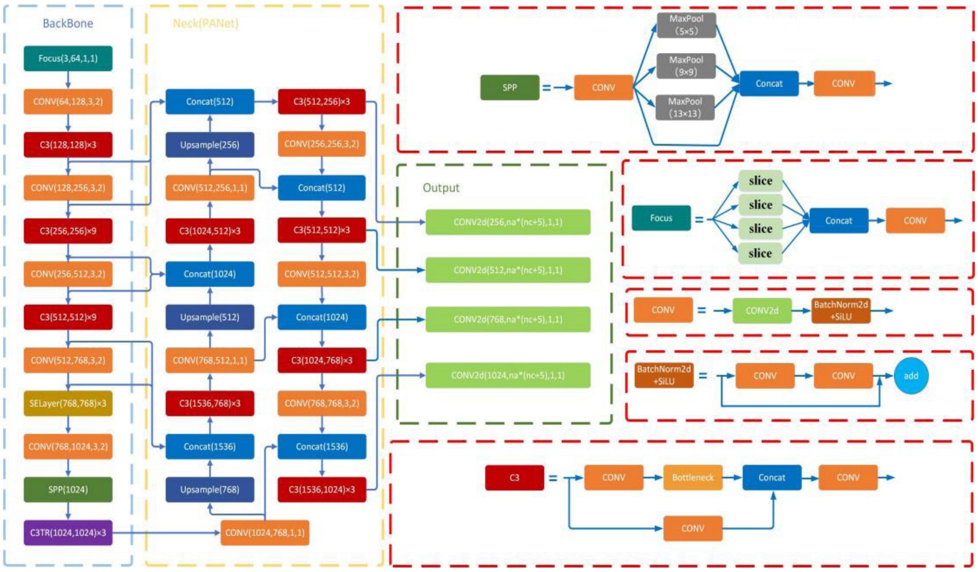


Fig. 4. Proposed target detection network

The Transformer can effectively extract global features through encoder and decoder. The encoder is composed of a multi-Head self-attention module and a feed-forward neural network, where the positional encoding solves the compression loss generated by the CNN network [19]. The decoder is constructed by various attention mechanisms, which improves the detector's efficiency for vehicle target detection. The structure of the Transformer is shown in Fig. 5.

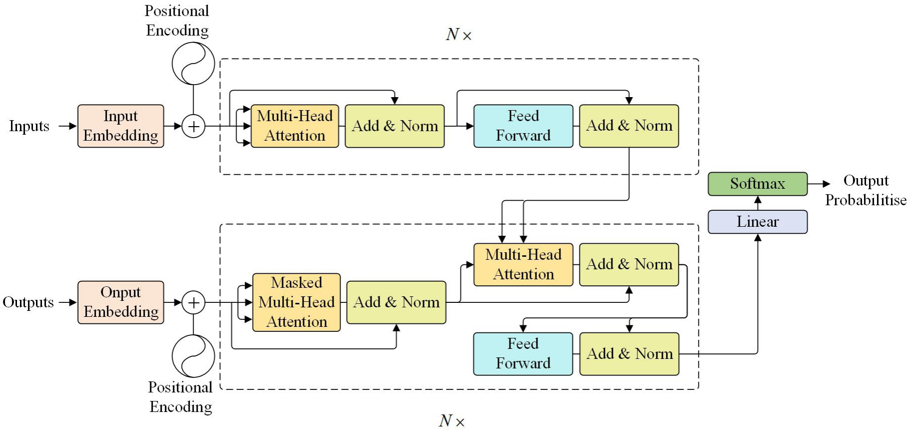


Fig. 5. Structure of Transformer

The low-level features and high-level features combine by the Transformer, which also links the features of different receptive fields of the spatial pyramid. However, not all the features extracted by Transformer are helpful. SE attention mechanism can selectively enhance useful features and suppress useless features through self-learning. As shown in Fig. 6, the current feature map is globally pooled, and weight is generated by self-learning through fully connected (FC) and activation operations. The weight is used to weight with the feature map and added to the original input simultaneously. In this way, the enhancement of valuable features and the suppression of useless features are achieved [20].

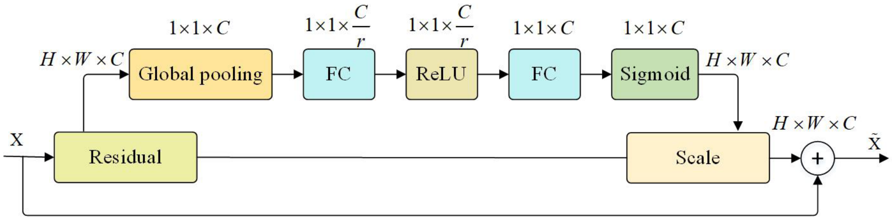


Fig. 6. SE attention mechanism

3.3 Target Tracking

The Deep SORT algorithm with good real-time and accuracy is adopted for better tracking of the non-motor vehicles in videos which have the characteristics of short stay time, large number, and small size. Deep SORT assumes the Tracking by Detection (TBD) strategy, takes the target detection result as the input, and estimates the object's trajectory through the Kalman filter combined with the target information extracted by the detector. The Hungarian algorithm is used to allocate the detection frame according to the location and appearance factors to achieve accurate tracking of the target, which guarantees the detection accuracy of violations [21]. The principle of the Deep Sort is shown in Fig. 7.

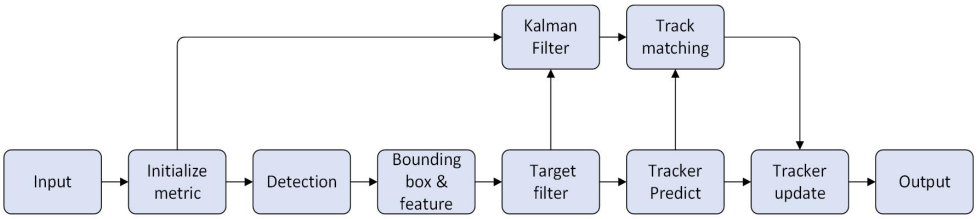


Fig. 7. Deep Sort algorithm

Considering that the actual application scenario environment is complex and changeable, primarily the phenomenon of multiple targets overlapping each other due to two cars being parallel and multi-car staggered occlusion, it is effortless to cause the ID assigned by the tracker to change. These phenomena will easily lead to missed detections and false detections in detection methods based on computer vision. To solve this problem, the motion trajectory and features of the targets are combined as a correlation metric, which avoids the tracking failure problem caused by the reappearance of the target after disappearing. The two types of features are correlated through the convolutional neural network. Occlusion or rapid movement causes the Mahalanobis distance used in Deep SORT not to be well-matched, so the minimum cosine distance is adopted. The minimum cosine distance between the i-th prediction result and the j-th detection result, as shown in Formula (1)

$$d^{(2)}(i,j) = \min\left\{1 - r_j^T r_k^{(i)} \mid r_k^{(i)} \in R_i\right\}, \quad \textbf{(1)}$$

Where r_j Represents each test result d_j calculate the appearance feature descriptor of the target. Set up $\left\|r_j\right\|$ =1，set a feature warehouse $R_i = \left\{r_k^{(i)}\right\}_{k=1}^{L_k}$ for each tracking track K and save nearly 100 feature descriptors associated with the results, L_k = 100.

3.4 Judgment of Violation

Common traffic violations of the non-motor vehicles include the driver not wearing a safety Helmet, illegally carrying people, and driving in reverse. To fill the gap in the supervision of intelligent transportation for non-motor vehicles is not comprehensive at present, this paper focuses on detecting the above violations. Frame by frame, the target features and motion trajectory extracted by detector and tracker are interacted with characteristics of non-motor vehicle violation to deduce violation behaviors.

The judgment of the behavior of not wearing a safety Helmet mainly depends on the features of the non-motorized vehicle and the features of the driver's Head. First, the detector is used to detect whether there are Two-wheel and Head targets in each frame of the video, and the target information is retained by the tracker. Then judge whether the Head position is in the bounding box of the Two-wheel and above the center point coordinates of the bounding box. If the Head position meets the above conditions, it is judged that the driver is not wearing safety a Helmet, as shown in Formula (2).

$$\begin{cases} L(X_{two}) \le (CHead) \le R(X_{two}) \\ L(Y_{two}) \le (CHead) \le R(Y_{two}) \end{cases}, \tag{2}$$

where L denotes the upper left corner of the Bounding box, X denotes the X-axis coordinate value, two denotes a Two-wheeled non-motorized vehicle, C denotes the center point, *Head* represents the driver's Head coordinates, and R denotes the upper right corner.

Similarly, according to the number of Heads targets or Helmets targets in the bounding box of Two-wheel in each frame to judge if the drivers are illegally carrying people, as shown in Formula (3).

$$\begin{cases} R(X_{two}) \ge (CHead) \ge L(X_{two}) \\ R(Y_{two}) \ge (CHead) \ge L(Y_{two}) \\ R(X_{two}) \ge (CHelmet) \ge L(X_{two}) \\ R(Y_{two}) \ge (CHelmet) \ge L(Y_{two}) \end{cases}, \tag{3}$$

where *Helmet* denotes the detected target is the Helmet.

The judgment of the non-motorized vehicle driving in the reverse direction mainly relies on the motion trajectory features extracted by the tracker. The first step is to determine whether the ID of the bounding box assigned by the tracker to non-motor vehicles is consistent in the current frame and subsequent frames. If the ID of two adjacent frames is identical, it is judged to be the same non-motor vehicle and continues to track the position information changes further. When the transformation of the bounding box in the current frame and subsequent frames does not meet the standard trajectory, it is determined that the non-motor vehicle has reverse driving behavior, as shown in Formula (4).

$$\begin{cases} ID_i = ID_{i+1} \\ y_i - y_{i+1} > 0 \end{cases}, \tag{4}$$

where ID denotes the id number of the non-motor vehicle, i denotes the current frame number, and y represents the coordinate of the y-axis of the non-motor vehicle.

4 Experimental Results and Analysis

In this section, firstly, experimental environment and evaluation indicator are introduced. Secondly, the experimental results are analyzed in detail to prove the effectiveness of the proposed algorithm.

4.1 Experimental Environment and Evaluation Indicator

To verify the effectiveness of the proposed algorithm, comparative experiments are performed on the new established database. And the database is divided into training set and validation set according to the ratio of 7:3. During training, the input size is set to 640×640, the batch is set to 8, and learning rate is 0.01.

The experimental environment is configured by a computer running Windows 10 64 bit, an AMD Ryzen7 5800X CPU, and an NVIDIA GeForce RTX3090 GPU. The proposed algorithm is implemented based on Pytorch1.7, cuda11.1, and CUDNN8.0.4.

Precision, Recall, Speed, Average Precious (AP), and mean Average Precious (mAP) are used as the evaluation indicators for target detection. In contrast, Multi-Object Tracking Accuracy (MOTA), Multi-Object Tracking Precision (MOTP), ID Switch are used as target tracking evaluation indicators. Speed indicates the detection speed, AP indicates the detection accuracy of a single class, mAP indicates the average detection accuracy, and ID Switch indicates the number of target label changes. In this paper, manual counting and comparative experiments are used to detect violations. There are relatively few samples of carrying people, which leads to a relatively high rate of missed detection of the experimental results. Setting a threshold for a single intersection during reverse driving detection dramatically reduces the missed detection rate, leading to a slight increase in the probability of false detection. Therefore, it is recommended to set different thresholds for different intersections.

$$Precision = \frac{TP}{TP + FP}, \quad \textbf{(5)}$$

where true positive (TP) is the number of positive samples correctly identified by the algorithm as positive samples, and false positives (FP) is the number of negative samples incorrectly identified by the algorithm as positive samples. Therefore, higher values of precision and recall mean better performance of the algorithm.

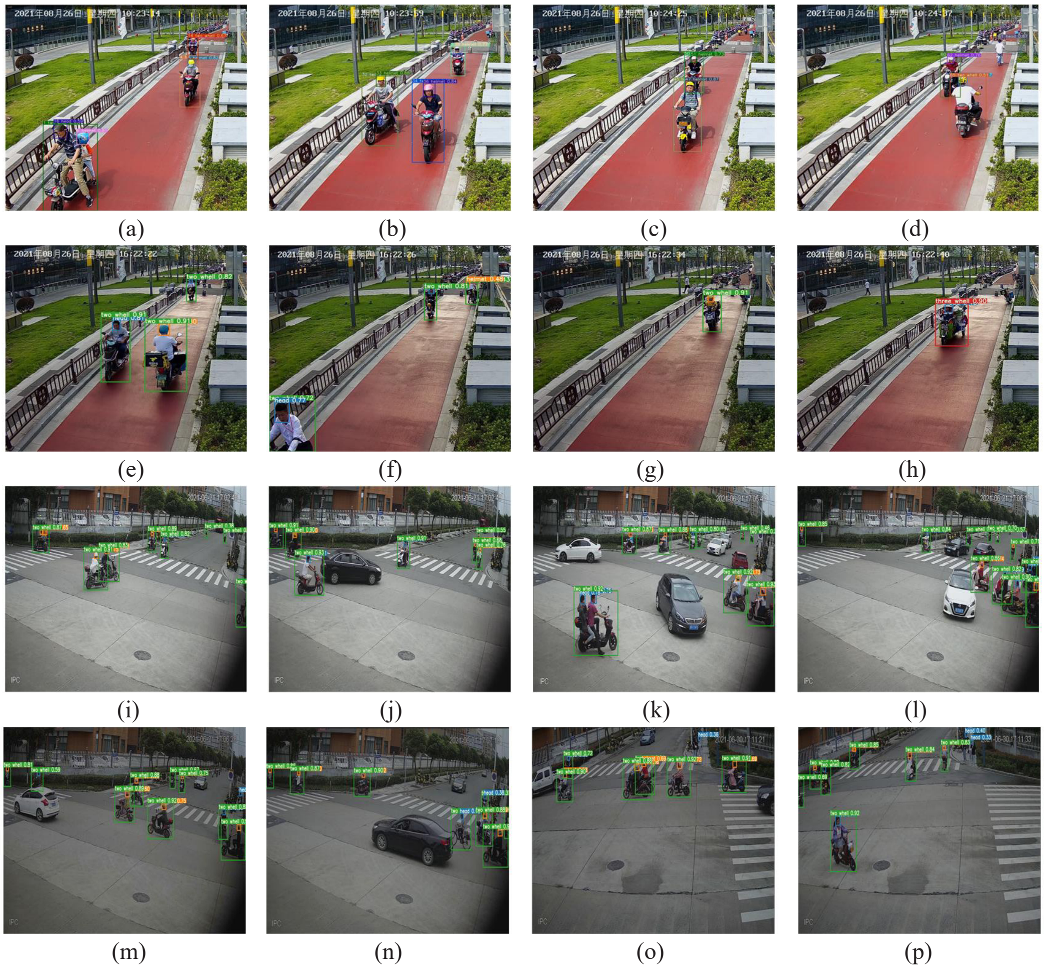
4.2 Experimental Results and Comparison

To verify the effectiveness of the proposed algorithm, which is compared with the existing advanced algorithms. And the experimental results are shown in Table. 1. Through comparison, it can be found that the proposed algorithm improves the performance of feature extraction by integrating the attention mechanism. By introducing Transformer, the positioning ability of the model for small targets is enhanced, and the SE attention mechanism is used to make the model pay more attention to the targets that need to be detected. Through this strategy, the influence of the accuracy of background changes is reduced and the generalization ability of the model is improved. Specifically, it can be found that the detection accuracy of helmet and head, which belong to small targets, are increased by 1.1% and 2.6% respectively compared with YOLOv5. And the detection accuracy of non-motor vehicles is increased by 2.7%. The mAP of the proposed algorithm reached 92.2%, which is an increase of 1.6% compared to YOLOv5. AP of each category has also been improved. Overall, the proposed algorithm has a significant improvement over the existing representative algorithms in dealing with the problems faced in this paper.

Table 1. Comparison of experimental results

Algorithms	Recall	AP				mAP
		Helmet	Head	Two-wheeler	Three-wheeler	
SSD	0.707	0.629	0.557	0.768	0.754	0.677
Fast R-CNN	0.756	0.705	0.644	0.825	0.849	0.756
YOLOv4	0.883	0.834	0.791	0.946	0.969	0.885
YOLOv5	0.872	0.915	0.791	0.969	0.948	0.906
Proposed	0.885	0.926	0.817	0.968	0.976	0.922

To test the generalization ability of the proposed algorithm in complex traffic environment, five videos of real traffic scene collected by road surveillance camera are tested, including intersections and non-motor lane in different shading scenes, as shown in Fig. 8. Fig. 8(a) to Fig. 8(d) show a multi-target detection task in a strong light environment; Fig. 8(e) to Fig. 8(h) show a multi-target detection task in a low light environment; Fig. 8(i) to Fig. 8(l) show a multi-target detection task under a complex background, including the interference of motor vehicles and the existence of occlusion. The difference between Fig. 8(m) to Fig. 8(p) and Fig. 8(i) to Fig.(l) show a low light environment. The detection results show that, the detector combined with attention mechanism can accurately locate and recognize helmet, head and non-motor vehicle targets in complex scenes, and the model trained with the database can effectively distinguish non-motor vehicles in progress from non-motor vehicles parked on the roadside, hence the proposed algorithm has good performance for single target and multiple targets in different scenes.



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

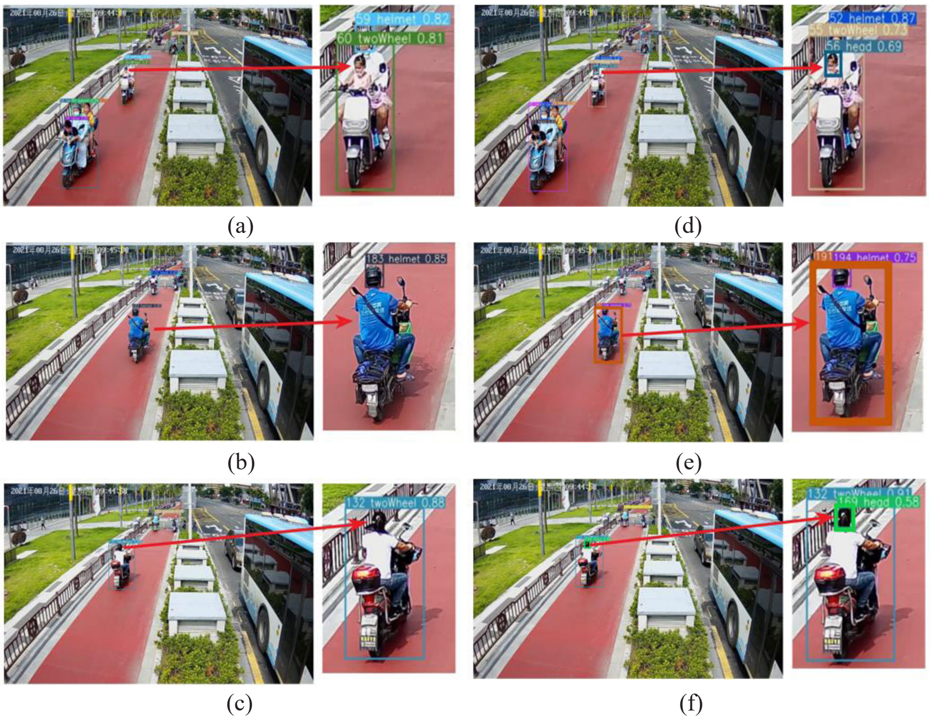
(m) (n) (o) (p)

Fig. 8. Detection results
(a) (b) (c) (d) is bright environment non-motor lane, (e) (f) (g) (h) is dark environment non-motor lane,
(i) (j) (k) (l) is bright environment intersection, (m) (n) (o) (p) is dark environment intersection

In addition, a tracking test is carried out in the MOT16 database [22], and the experimental results are shown in Table. 2, where ID Swich is used to assign an ID to each target and record their trajectories. ID swich is the number of ID changes in the tracking, and the larger it is, the more unstable the tracking is. Compared with the existing representative algorithms, the ID switch of the proposed algorithm is reduced to 28 and has a higher tracking speed. The experimental results show that introducing the Transformer network and fusing the low-level and high-level features lead to improve feature extraction performance of global information, and the attention mechanism can further increase the recognition ability of the algorithm for moving helmet, head and non-motor vehicle targets in the real scene. With the help of the detection algorithm, searching in the global view can obtain better tracking results, an ID will assign to each target, then the nearest neighbor matching is carried out through the appearance characteristics of the targets, as shown in Fig. 9. And experimental results show that the proposed algorithm achieves better tracking performance. In addition, when the tracking target has different degrees of occlusion, including partial occlusion and large-area occlusion, the proposed algorithm can still identify and track accurately and efficiently, as shown in Fig. 10.

Table 2. Comparison of tracking results

Algorithms	ID Swith	Speed (Hz)
SORT	75	100
Deep SORT	39	91
Proposed	28	83

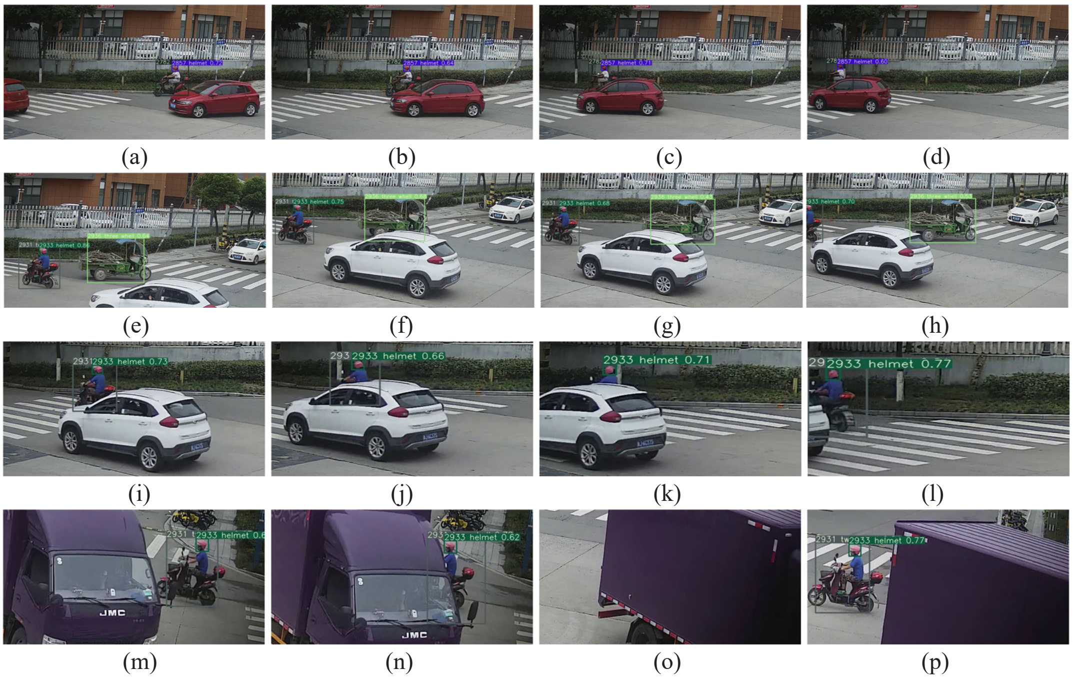


(a) (d)

(b) (e)

(c) (f)

Fig. 9. Comparison of experimental results
(a)(b)(c) is Yolov5+Deep SORT, (d)(e)(f) is the proposed



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

(m) (n) (o) (p)

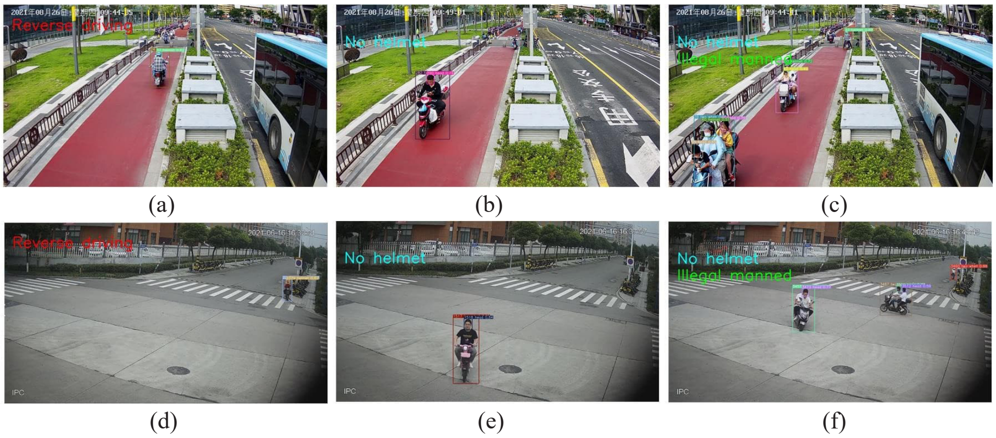
Fig. 10. Detection and tracking results with different degrees of occlusion

Considering the complex and changeable traffic conditions, this paper tests five video streams. The comparison of manual statistics and the proposed algorithm statistics in this paper is shown in the following table. Are shown in Table. 3. The statistical results prove that the missed detection rate and false detection rate of the proposed algorithm for the three types of violations are all at low values, among them, the target detection network combined with the attention mechanism can accurately locate and identify the target, extract the location information of the target, and accurately infer the violations of not wearing a helmet and illegal manned behavior through logical judgment. The behavior of driving in reverse is difficult to judge by target detection network, so using network combined detection and tracking network, and through the target motion trajectory generated by the tracker to judge the violation behavior, and the detector's good feature extraction ability effectively avoids the problem of ID Switch in the tracking and reduces the number of alarms for the same violation event, and considering the behavior of driving in reverse accounts for the most. So, Equation (1) is defined to reduce the missed detection rate.

In order to further verify the ability of the proposed algorithm to detect violations in different scenes, two different videos of actual traffic scenes have been tested. The experimental results show that the feature fusion enhanced network by fusion of shallow and deep features, can identify targets with different angles and sizes in non-motor lane and intersection, then through logical judgment, the algorithm can accurately identify violations in video streams, and issue corresponding warnings according to the identified violations. The results are shown in Fig. 11.

Table. 3. Detection results of violation

Violations	Manual statistics	Algorithm statistics
Not wear safety helmet	28	25
Illegally carry people	15	13
Drive in reverse	51	47



(a) (b) (c)

(d) (e) (f)

Fig. 11. Warning of violations
(a) (d) driving in reverse, (b) (c) not wearing helmet, (c) (f) illegally carrying people

5 Conclusion

In this paper, a non-motor vehicle violation detection algorithm with attention mechanism is proposed, which can accurately identify the violation in real-time. Firstly, in order to ensure the applicability of the algorithm, a real traffic scene database collected by surveillance cameras is established, and reasonable annotation methods are used to effectively distinguish running and parked non-motor vehicles. Secondly a detection network combined with a Transformer is constructed, which improves the connection of contextual features and performance of global features extraction. In order to optimize the extraction of local features, the Squeeze-and-Excitation channel attention mechanism is integrated into the network. Global and local joint feature extraction strengthen the detection performance of the network for small targets. Subsequently, the fusion of shallow and deep features avoids the ID Switch problem when Deep SORT performs target tracking. Finally, the violation judgment module set for the actual scene realizes the real-time detection of non-motor vehicle violations in the video stream. Real detection and data statistics show that the proposed algorithm can accurately detect traffic violations such as non-motor vehicle drivers not wearing a Helmet, illegally carrying people and driving in reverse in high-density scenarios, which has good practical value. In the future, deep learning still has great development space and application potential in smart transportation. It can greatly save labor costs, and easy to deploy. In future works, we look forward to realize multi-modal detection and tracking by introducing depth information collected by laser lidar or binocular camera to achieve better results, so as to contribute to the development of intelligent transportation system.

6 Acknowledgement

This work was supported by the National Natural Science Foundation of China (61671412); Natural Science Foundation of Zhejiang Province (LY21F010014, LY19F010002); Natural Science Foundation of Ningbo, China (2018A610053, 202003N4323); General Scientific Research Project of Zhejiang Education Department (Y201941122); Ningbo Municipal Projects for Leading and Top Talents (NBLJ201801006); Fundamental Research Funds for Zhejiang Provincial Colleges and Universities; School-level Research and Innovation Team of Zhejiang Wanli University; Science and Technology Innovation 2025 Major Project (2022Z076).

References

[1] R. Girshick, Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation, in: Proc. 2014 IEEE Confer-ence on Computer Vision and Pattern Recognition, 2014.
[2] R. Girshick, Fast R-CNN, in: Proc. 2015 IEEE International Conference on Computer Vision, 2015.
[3] S.-Q. Ren, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, IEEE Transactions on Pat-tern Analysis and Machine Intelligence 39(2015) 1137-1149.
[4] J. Uijlings, Selective Search for Object Recognition. International Journal of Computer Vision 104(2013) 154-171.
[5] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, SSD: Single Shot Multi Box Detector, in: Proc. 2016 Europe Conference on Computer Vision, 2016.
[6] C. Fu, DSSD: Deconvolutional Single Shot Detector. <https://arxiv.org/abs/1701.06659>, 2017 (accessed 23.01.2017).
[7] J. Redmon, You Only Look Once: Unified, Real-Time Object Detection, in: Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016.
[8] J. Redmon, A. Farhadi, YOLO9000: Better, Faster, Stronger, in: Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017.
[9] J. Redmon, A. Farhadi, YOLOv 3: An Incremental Improvement. <https://arxiv.org/abs/1804.02767>, 2018 (accessed 08.04.2018)
[10] B. Alexey, YOLOv4: Optimal Speed and Accuracy of Object Detection. <https://arxiv.org/abs/2004.10934>, 2020 (accessed 23.04.2020).
[11] T.-Y. Lin, Feature Pyramid Networks for Object Detection, in: Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017.
[12] S. Liu, Path Aggregation Network for Instance Segmentation, in: Proc. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.
[13] A. Bewley, Simple online and real-time tracking, in: Proc. 2016 IEEE International Conference on Image Processing, 2016.
[14] J. Feng, Non-motor vehicle illegal behavior discrimination and license plate detection based on real-time video, in: Proc. Journal of Physics: Conference Series, IOP Publishing, 2020.
[15] R. V. Silva, Automatic detection of motorcyclists without helmet, in: Proc. 2013 XXXIX Latin American Computing Confer-ence, 2013.
[16] C. Wen, The safety helmet detection for ATM's surveillance system via the modified Hough transform, in: Proc. IEEE 37th Annual 2003 International Carnahan Conference on Security Technology, 2003.
[17] J. Mistry, An automatic detection of helmeted and non-helmeted motorcyclist with license plate extraction using convolutional neural network, in: Proc. 2017 Seventh International Conference on Image Processing Theory, Tools and Applications, 2017.
[18] A. Hirota, Classifying Helmeted and Non-helmeted Motorcyclists, in: Proc. International Symposium on Neural Networks, 2017.
[19] A. Vaswani, Attention is All you Need. <https://arxiv.org/abs/1706.03762>. 2017 (accessed 12.06.2017).
[20] J. Hu, Squeeze-and-Excitation Networks, IEEE Transactions on Pattern Analysis and Machine Intelligence 42(2020) 2011-2023.
[21] N. Wojke, Simple online and real-time tracking with a deep association metric, in: Proc. 2017 IEEE International Conference on Image Processing, 2017.
[22] L. Laura, MOTChallenge 2015: Towards a Benchmark for Multi-Target Tracking. <https://arxiv.org/abs/1504.01942>, 2015 (accessed 08.04.2015).