International Journal for Housing Science and Its Applications
Publish August 6, 2025. Volume 46, Issue 3 Pages 5159-5167

https://doi.org/10.70517/ijhsa463443

Real-Time Target Detection Algorithm and Event Analysis for Complex Traffic Scenes Based on Multimodal Data Fusion

Yukang Zou[1,*] and Xianjun Tan[2]
[1] School of Traffic and Transportation Engineering, Central South University, Changsha, Hunan, 410075, China
[2] Institute of Rock and Soil Mechanics, Chinese Academy of Sciences, Wuhan, Hubei, 430071, China
Corresponding authors: (e-mail: 19307411856@163.com).

Abstract Aiming at the problems of low target detection accuracy, poor real-time multi-target tracking and difficulty in recognizing small targets in complex traffic scenes, this paper proposes a real-time target detection algorithm based on improved YOLOv5s. A directed graph scene model containing environment features and object features is constructed, and the Marginalized Kernel algorithm is used to enhance the dynamic environment sensing ability. Improve the model architecture of YOLOv5s and optimize the feature extraction with the help of MHSARM. Enhance the spatial localization by combining CoordConv, and realize the joint learning of target detection and epigenetic features based on JDE paradigm. Experimental results show that on the TT100K dataset, the model in this paper outperforms all comparative models, with a 22.92% improvement in mAP@0.5 compared to the YOLOv5s baseline model, achieving an accuracy of 86.17%, and also demonstrating the best detection performance on the BDD100K dataset. The improved model performed best in terms of AP@0.5 accuracy in ablation experiments, achieving a mAP value of 80.24% in validation across six types of real traffic scenarios.

Index Terms target detection, Marginalized Kernel algorithm, YOLOv5s model, MHSARM, JDE paradigm

I. Introduction

With the continuous advancement of global urbanization, the number and categories of traffic participants are increasing, and in the real traffic environment often presents the characteristics of high dynamics and multi-category mixing, which makes the traffic safety problem more and more prominent [1], [2]. As a new stage in the development of intelligent transportation, the vehicle-road cooperative system is able to sense the information of traffic participants in real time and accurately, and share the data fusion with all the traffic subjects, which can effectively solve the problems faced by the current urban transportation [3]-[6]. Real-time and accurate detection of traffic participants by road monitoring cameras can provide effective data protection for the vehicle-road cooperative system, which has evolved into one of the popular research directions in the field of intelligent transportation [7]-[9].

Target detection algorithms play a key role in the vehicle-road cooperative system, which helps to prevent collisions and improve road safety by monitoring and recognizing other vehicles, pedestrians, bicycles and other targets in real time [10], [11]. With the acceleration of urbanization, traffic flow and complexity are increasing, and the demand for intelligent transportation systems and urban planning is becoming more and more urgent [12], [13]. Traffic target detection technology can provide real-time and accurate traffic data for urban planning and support more scientific urban traffic management [14], [15]. At the same time, in order to promote the promotion and application of autonomous driving technology, governments have gradually introduced relevant regulations and policies [16]. This further promotes the research on traffic target detection algorithms to meet the requirements of regulations on safety and accuracy.

In this paper, we first construct a multimodal scene model containing the environment semantic segmentation results and dynamic object attributes, and represent the global and local information in the scene through directed graphs. The Marginalized Kernel algorithm is used to analyze the similarity between scene graphs and capture the dynamic evolution of the traffic scene. The MHSARM module is introduced to optimize the feature extraction process of Yolov5s model, and CoordConv is used to replace the traditional convolution operation. Combined with the JDE paradigm, the network branch of epigenetic feature extraction is added to the optimized YOLOv5s head network to achieve real-time multi-target tracking of complex traffic scenes. The model is trained to generate anchor frames relying on K-means combined with prior knowledge. The effectiveness of the improved scheme is verified through comparison experiments and ablation experiments. Six scenarios are selected to carry out practical validation to examine the application effect of this paper's model in real scenarios.

II. Design of real-time target detection algorithm for complex traffic scenes based on multimodal data fusion

With the development of intelligent transportation systems, real-time target detection and event analysis in complex traffic scenarios have become the core challenges in the fields of autonomous driving and video surveillance. In real scenarios, large differences in target scales, complex background interference and dynamic environment changes lead to the traditional detection algorithms facing problems such as accuracy degradation, high leakage detection rate and limited computational resources. Existing methods mostly focus on the feature extraction of single modal data (e.g., RGB images), ignoring the synergistic analysis of environmental semantics and dynamic features of objects, and it is difficult to balance the detection accuracy and real-time demand.

II. A. *Scene posture analysis based on Marginalized Kernel algorithm*

II. A. 1) Multimodal scene modeling

In order to be able to describe the information in the scene from both the macro overall environment perspective and the micro individual object perspective, the scene model designed in this paper is divided into two parts, environment features and object features. The environment features are the semantic segmentation results of the images from the front camera. The semantic segmentation results of images can provide holistic pixel-level image classification results, which can help the subsequent algorithms understand the static environmental information in the scene. Similar to images, semantic segmentation results for images are tensors with structure $C \times H \times W$.

Object features, on the other hand, are descriptions of the specific properties of each movable object in the scene to help subsequent algorithms understand the dynamic object information in the scene. In order to maintain the indeterminacy of the number of objects in different scenes, the scene model constructed in this paper is shown in Fig. 1. In this paper, we use an indeterminate-length 2nd-order tensor, $(N+1) \times f_n$, as the object features in the scene model to describe the object information in the scene. Where N denotes the number of movable objects in the scene, 1 denotes the current vehicle itself, and f_n denotes the features of each object.

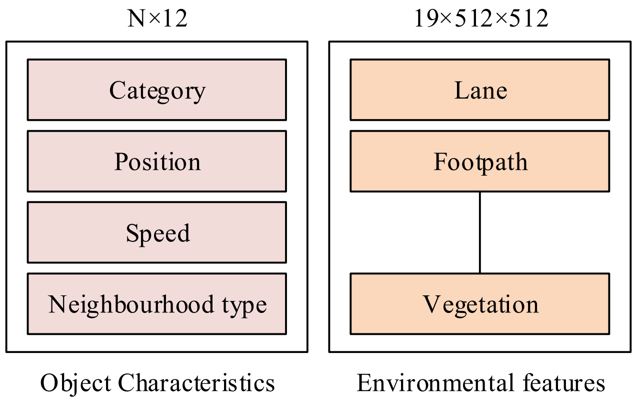

Figure 1: Scene model

II. A. 2) Marginalized Kernel Algorithm

Marginalized kernel algorithms are a class of graph kernel algorithms based on directed graphs that can be used to analyze the similarity between different directed graphs.Marginalized kernel algorithms use multiple paths obtained by sampling in a directed graph using random walks. For any path h of graph G, it is sampled by choosing any initial point from all nodes with initial probability $p_s(h_1)$ of $1/n$. At the ith step of sampling, the probability of shifting to the proximity point is $p_t(h_i \mid h_{i-1})$ and there is a probabilistic end of $p_q(h_{i-1})$. The relationship between the two is shown in equation (1).

$$\sum_{j=1}^{|G|} p_t(j \mid i) + p_q(i) = 1 \tag{1}$$

In this paper the size of $p_q(h_{i-1})$ is set to 0.1. Then the probability of sampling a path h from a graph G is obtained by calculating equation (2).

$$p(h \mid G) = p_s(h_1) \prod_{i=2}^{\ell} p_t(h_i \mid h_{i-1}) p_q(h_\ell) \tag{2}$$

For any pair of paths sampled from a graph G vs. a graph G', use Eq. (3) to compute its kernel function.

$$K_z(z,z') = \begin{cases} 0 & (\ell \neq \ell') \\ K\left(v_{h_1}, v'_{h'_1}\right)\prod_{i=2}^{\ell} K\left(e_{h_{i-1},h_i}, e'_{h'_{i-1}} h_i'\right) \cdot K\left(v_{h_i}, v'_{h'_i}\right) & (\ell = \ell') \end{cases} \tag{3}$$

Equation (3) with $z=(G,h)$.

In this paper, the kernel function between nodes is $K(v,v')=\frac{1}{2}\delta(v=v')$, and the kernel function between edges is $K(e,e')=\exp\left(-\|e-e'\|^2 \big/ 2\sigma^2\right)$ Thus, the similarity between G and G' can be calculated by Eq. (4).

$$\begin{aligned} K(G,G') = {} & \sum_{\ell=1}^{\infty}\sum_{h}\sum_{h'} p_s(h_1)\prod_{i=2}^{\ell} p_t(h_i \mid h_{i-1})\, p_q(h_l) \\ & p_s'(h_1')\prod_{j=2}^{\ell} p_t'(h_j' \mid h_{j-1}')\, p_q'(h_\ell') \\ & K\left(v_{h_1}, v'_{h'_1}\right)\prod_{k=2}^{\ell} K\left(e_{h_{k-1}h_k}, e'_{h'_{k-1},h'_k}\right) K\left(v_{h_k}, v'_{h'_k}\right) \end{aligned} \tag{4}$$

II. B.Traffic target detection based on improved Yolov5s algorithm

II. B. 1) Improvement of the overall architecture of Yolov5s

Taking Yolov5s as the basic framework, the improvement consists of 2 parts: (1) Replacing the C3 module in Backbone by utilizing the Multihead Self-Attention Mechanism Residual Module (MHSARM); and (2) Introducing CoordConv in the feature fusion region of Yolov5s.MHSARM fully combines the excellent feature extraction capability of residual structure and the MHSA's excellent attention ability, in addition to ensuring excellent feature learning ability, it can assign weights according to the feature information, so as to suppress the background interference and strengthen the target information. In the Yolov5s feature extraction region, the SPPF structure is the last module in the feature extraction region, which serves to convert the feature map of arbitrary size into a fixed-size feature vector, which contains strong semantic information of the image, reduces the interference brought by the complex road to the SPPF structure, and can effectively improve the detection efficiency of the algorithm, so the C3 module for obtaining the SPPF in the basic algorithm is used as the MHSARM. MHSARM to replace it.

II. B. 2) MHSA-Bottlenck

The principle of MHSA-Bottleneck is to access the multi-head self-attention mechanism after 3*3 convolution in Bottleneck, and its structure is shown in Figure 2. Considering the balance between real-time and detection accuracy, this study adopts a 4-head self-attention mechanism, whose advantages are reflected in the following three points: (1) to ensure that the model will not increase the number of parameters too much, resulting in real-time affected; (2) will not lead to a decline in the feature learning ability because of the deepening of the number of layers of the self-attention mechanism; (3) Combine Bottleneck's excellent feature learning capability with MHSA's excellent anti-interference capability to provide better detection performance in complex traffic scenes.

$$head_i = Attention\left(QW_i^Q, QW_i^K, QW_i^V\right) \tag{5}$$

$$MultiHead(Q,K,V) = Concat\left(head_1, \cdots, head_h\right) \tag{6}$$

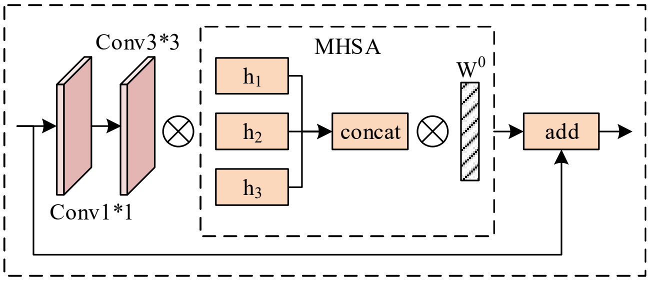

Figure 2: MHSA-Bottleneck structure

II. B. 3) MHSARM

The C3 module is an important part of the Yolov5s backbone and feature fusion network, which combines the advantages of depth-separable convolution and null convolution.C3 consists of three convolutional layers as well as a Bottleneck module, in which the Conv1 step is 2, whose purpose is to increase the sensory field of the network, so that the network can better retain the global information.The Conv2 and Conv3 steps are 1 , whose purpose is to preserve the spatial resolution of the feature map so as to better retain the local feature information of the target. However, the C3 module lacks the ability to focus on the target, which leads to its limited ability to extract features in complex backgrounds, especially in scenes where feature information is easily lost (strong light, backlight, dense) extracting features is more difficult, and there are low detection accuracy, leakage, and misdetection phenomena.

The structure of MHSARM is shown in Fig. 3, the principle of this module is that the Bottleneck module in C3 is replaced by MHSA-Bottleneck, which not only inherits the advantages of MHSA-Bottleneck, but also combines the feature extraction ability of C3 at different scales.MHSARM, while possessing excellent feature extraction ability, can also be used to assign weighting coefficients to features according to feature correlation. MHSARM can also assign weight coefficients to features based on feature relevance, assigning larger weights to information that needs to be emphasized and smaller weights to irrelevant background information. In the field of automatic driving target detection, MHSARM focuses more attention on traffic targets, so as to achieve the purpose of weakening background interference and strengthening target feature information. After experiments, it is proved that MHSARM can effectively improve the detection performance in the face of complex scenes of various types (strong light, backlight, dense).

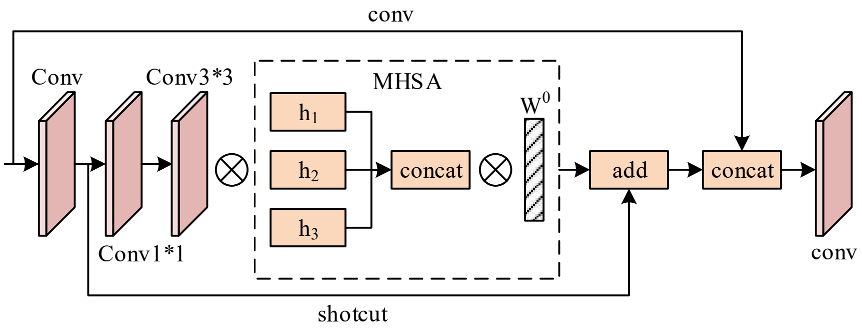

Figure 3: MHSARM structure

II. C.Multi-target tracking based on YOLO and JDE

The JDE paradigm proposes the design idea of joint target detection and epigenetic features, which solves two tasks of target detection and epigenetic feature extraction simultaneously through a shared backbone network, reduces redundant computation, and has good real-time performance. Therefore, based on the JDE paradigm, this paper proposes a real-time multi-target tracking model by adding an additional network branch for epigenetic feature extraction in the optimized YOLOv5s head network. Among them, the feature maps output from the added epigenetic feature extraction network branch contain the feature vectors of the targets. When determining the feature vector of the target, this paper will stitch the feature maps output from the target detection branch and the apparent feature extraction branch along the channel direction, correlate the detection frame with the feature vector, and determine the feature vector of the target according to the detection frame.

When training the epigenetic feature extraction network branch, this paper converts the epigenetic feature extraction task into a classification task, and takes the target ID as the category of the target feature vector. According to the positive sample setting method of YOLOv5s, there are 3 positive sample anchors for each labeling box, and each anchor will correspond to a feature vector, so each target will have 3 feature vectors.

Each target has 3 annotation boxes, and the center of each annotation box is used to determine the anchor points for positive samples. The coordinates of the 3 anchor points on the feature map can determine the corresponding 3 feature vectors. Each annotation box can well surround the target, so the corresponding feature vectors can also effectively express the appearance features of the target, and the category of each feature vector corresponds to the target ID. In the end, the task of extracting appearance features can be transformed into a classification task.

The classification task of epigenetic feature extraction is then solved by designing a fully-connected network layer, where the feature vectors are used as inputs and the target IDs are used as category labels, and the fully-connected layer predicts the probability that the feature vectors belong to each target ID. In this case, the feature vectors are normalized before being input to the fully connected layer, which is calculated as shown in Eqs. (7) and (8).

$$V_{in} = [e_1, e_2, \cdots, e_k] \tag{7}$$

$$V_{out} = \frac{V_{in}}{\sqrt{\sum_{i=1}^{k}(e_i^2)}} \tag{8}$$

where V_{in} denotes the input feature vector, k denotes the dimension of the feature vector, and V_{out} denotes the normalized feature vector.

The mainstream cross-entropy loss function is used to calculate the classification loss, as shown in Equation (9).

$$Loss_{ce} = -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n} v_j^i \cdot logp_j^i \tag{9}$$

where m denotes the number of feature vectors, n denotes the number of target IDs (categories), v_j^i denotes the true value of the i th feature vector in the j th category, and p_j^i denotes the predicted value of the i th feature vector in the j th category.

After calculating the loss value by the cross-entropy loss function, the fully connected network layer and the target epigenetic feature extraction network branch are optimized by gradient descent method, so as to extract the discriminative feature vectors.

III. Analysis of the application effect of real-time target detection algorithms for complex traffic scenes

III. A. Model training

In order to verify the performance advantage of the model in this paper for target detection, the TT100K dataset and the BDD100K dataset are chosen to be used for experiments. For the TT100K dataset, by analyzing the data characteristics of different types of traffic signs in it, 50 categories containing more than 100 instances are selected to reduce the problem of sample discrepancy due to the large gap in the number of category instances.

The experiment will train the model weights from scratch for all the involved target detection models to minimize the influence of external factors and ensure fairness. The models are trained on two datasets using the SGD optimizer. The datasets will be divided into three parts in the ratio of 8:1:1, including the training set, validation set and test set. The hyperparameters used in the experiments were configured as follows: a Batchsize of 32, an initial learning rate of 0.01, a final learning rate of 0.01, a momentum of 0.937, and a weight decay of 0.0005. The loss function value is made to drop to a certain range by continuous iteration and training is stopped, the process focuses on the loss function and the average IOU value, the results of the visualization of the loss function and IOU changes are shown in Fig. 4.

From the figure, it can be seen that with the training process, the loss of the network gradually decreases when the number of iterations is 4000, the network loss no longer changes. At the same time, the IOU value gradually becomes larger with training and finally stabilizes to 1. Therefore, when choosing the final model, the model with 4000 iterations is an important cut-off point.

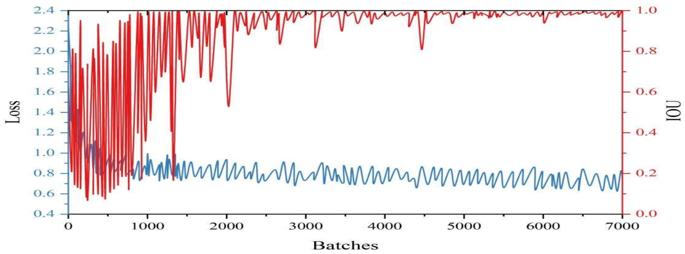

Figure 4: Visualization results of loss function and IOU change

Another important parameter involved in the training of the model is the “anchor frame”. The model generates corresponding anchor frames in each cell of each prediction layer according to the pre-setting as the initial prediction, and then generates the final prediction after the prediction results of the non-maximal value suppression and classification network. Therefore, the selection of anchor frames has a direct impact on the accuracy of the model. The anchor frame is set from the training data, and its size and proportion should correspond to the target size distribution in the data set to ensure that the anchor frame can accurately describe the size and proportion distribution of the target in the data set. In this paper, we use K-means combined with prior knowledge to generate anchor frames, and the specific steps are as follows:

(1) Give the position of the initial anchor frame in the scatter plot;

(2) Calculate the distances from all targets in the dataset to each anchor point separately, select the nearest point as the clustering point, and calculate the distances from all points to the clustering point and L at this time;

(3) Update the position of the cluster point and repeat the operation of (2) until the value of L no longer changes. The clustering results are shown in Fig. 5, and finally the anchor frames with four kinds of aspect ratios are clustered.

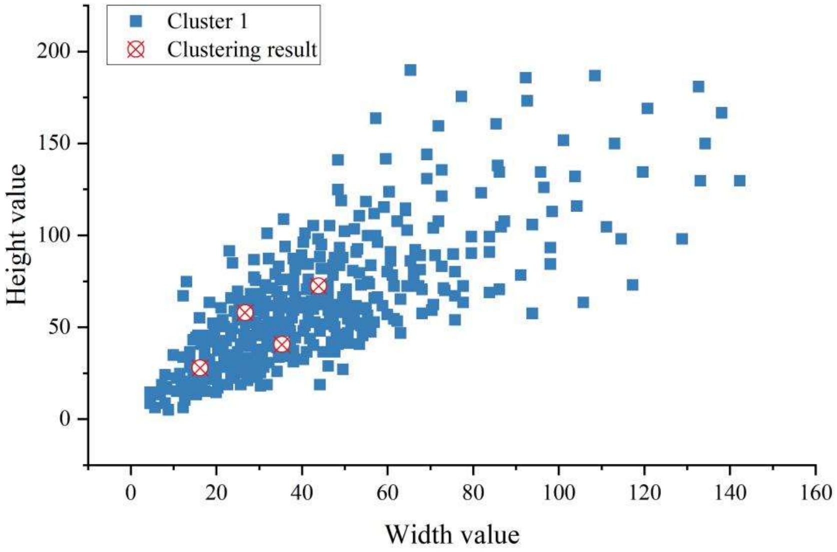

Figure 5: Clustering results

III. B. Experimental analysis

III. B. 1) Comparative experiments

The experimental results of different models on the TT100K dataset are shown in Table 1, which clearly displays the comparison. First of all, compared to the YOLOv5s baseline model, the model presented in this paper shows a significant improvement in performance, with mAP@0.5 increasing by 22.92%, reaching an accuracy of 86.17%. This indicates that the model effectively addresses the limitations of the YOLOv5s baseline model in small object detection. Moreover, the performance of this model far exceeds that of two-stage models, such as YOLOv5m and YOLOv7, which did not demonstrate good results in traffic scene data despite having a higher parameter count. Additionally, compared to some state-of-the-art object detection models like YOLOv7-Tiny and YOLOv8s, this model also outperforms them, achieving the best results. It is worth noting that the improved model only adds an extra 2.07M parameters and still maintains real-time detection capability. Among lightweight models with similar parameter counts, this model shows a clear advantage in detecting small traffic targets.

Table 1: Comparison of results on the TT100K dataset

Model	Backbone	mAP/%	Param/M	Speed/ms
Faster R-CNN	ResNet-50	52.53	40.63	27.36
Cascade R-CNN	ResNet-50	69.25	62.45	39.63
Sparse R-CNN	ResNet-50	62.42	98.42	36.36
RetinaNet	ResNet-50	42.64	39.77	28.48
YOLOv5s	CSP-Darknet53-C3	63.25	8.31	10.42
YOLOv5m	CSP-Darknet53-C3	74.21	30.24	13.11
YOLOv7	ELAN-Net	69.35	38.17	10.48
YOLOv3-Tiny	Darknet53	63.58	9.13	4.20
YOLOv4-Tiny	CSP-Darknet53	52.19	4.14	4.19
YOLOv7-Tiny	ELAN-Net	42.64	7.05	9.01
YOLO-MAXNet	SA-MobileNeXt	72.59	15.25	16.32
YOLOv8s	CSP-Darknet53-C2f	80.32	12.53	9.34
The proposed	CSP-Darknet53-C3	86.17	10.38	9.37

In order to further prove the effectiveness of the model in this paper, twelve more mainstream target detection algorithms were selected and experimented on the BDD100K dataset, and the resultant data are shown in Table 2. Since the BDD100K dataset contains a large number of multi-scale targets, the experiments focus on comparative evaluation for small-scale targets, so the detection results (AP) of different algorithms for the two categories of small targets (traffic signs and traffic signals) in this dataset, as well as the average detection accuracies (mAP@0.5) for all the categories, are specifically listed. According to the data in the table, it can be seen that the AP values of this paper's model for traffic signs and traffic signals reach 74.58% and 69.21%, respectively, which is a significant performance improvement compared to other detection models, and has reached the optimal results in small target detection. At the same time, the model in this paper still maintains a lightweight model parameter number and has the speed to be able to perform real-time detection, with the mAP value and parameter number of 80.08% and 10.31M, respectively.

Table 2: Comparison of results on the BDD100K dataset

Model	Backbone	Trafic Sign/%	Trafic Lignt/%	mAP/%	Param/M	Speed/ms
Faster R-CNN	ResNet-50	67.35	58.24	73.42	42.42	38.34
Cascade R-CNN	ResNet-50	68.24	57.33	74.24	62.35	46.22
Sparse R-CNN	ResNet-50	66.59	59.62	70.55	101.35	36.46
RetinaNet	ResNet-50	67.83	58.48	75.17	34.24	34.21
YOLOv5s	CSP-Darknet53-C3	69.32	53.12	74.28	8.03	8.43
YOLOv5m	CSP-Darknet53-C3	70.14	62.42	75.09	27.46	10.35
YOLOv3	Darknet53	72.49	65.23	76.27	63.53	10.22
YOLOv4	CSP-Darknet53	72.97	66.93	77.05	65.34	15.23
YOLOv3-Tiny	Darknet53	37.35	25.35	49.23	9.03	3.12
YOLOv4-Tiny	CSP-Darknet53	44.23	27.34	54.23	4.22	4.13
YOLOv7-Tiny	ELAN-Net	59.24	58.21	67.23	7.03	7.34
YOLO-MAXNet	SA-MobileNeXt	62.18	69.23	72.42	14.24	18.23
YOLOv8s	CSP-Darknet53-C2f	63.85	59.29	74.28	12.65	8.43
The proposed	CSP-Darknet53-C3	74.58	69.21	80.08	10.31	8.64

The extensive experimental data mentioned above show that the model in this paper is well suited for detecting small targets in the transportation domain. In addition, the performance advantages demonstrated on two different datasets further confirm the generality of this paper's model in solving the small target detection problem.

III. B. 2) Ablation experiments

In order to evaluate the contribution of each improvement component, a large number of ablation experiments were conducted on the TT100K dataset to verify the effectiveness of each of the proposed schemes. The different schemes and their corresponding experimental result data are shown in Table 3.

The first row of the table presents the experimental results of the YOLOv5s baseline model, with subsequent improvement plans based on adjustments to the YOLOv5s architecture. Plan (a) replaces the C3 module in the Backbone of the baseline model with MHSARM. Comparing the data in the table shows that after using MHSARM, the model's AP@0.5 increased from the baseline 63.25% to 75.23%, while the model's parameter count remained nearly unchanged. This indicates that the design of MHSARM significantly enhances the model's feature fusion capability without any cost, which can notably improve detection accuracy for small targets. Plan (b) demonstrates the performance of introducing CoordConv in the feature fusion area of YOLOv5s. Plan (c) shows the performance after adding a network branch for extracting apparent features. Plan (d) is a combination of plans (a) and (b), plan (e) is a combination of plans (a) and (c), plan (f) is a combination of plans (b) and (c), and plan (g) represents the experimental results of incorporating all design plans. Compared to the baseline model, the proposed model improves AP@0.5 accuracy by 22.92%, and the experimental results fully demonstrate that the architectural design of this model enhances performance.

Table 3: Ablation Experiment Results

Model	MHSARM	CoordConv	JDE	mAP/%	Param/M
YOLOv5s				63.25	8.31
(a)	√			75.23	8.31
(b)		√		69.42	9.02
(c)			√	73.52	9.98
(d)	√	√		82.91	9.15
(e)	√		√	84.87	10.24
(f)		√	√	82.66	10.38
(g)	√	√	√	86.17	10.38

III. C. Empirical analysis

In this paper, six scenarios are selected to launch the practical verification. Scene 1 is a sidewalk intersection with more dispersed between targets, Scene 2 is a congested road with serious occlusion between vehicles at dusk time, Scene 3 is a highway with large differences in target sizes, Scene 4 is a viaduct with a complex background and small target sizes, Scene 5 is a highway with a large traffic flow, and Scene 6 is an urban and rural road with small target sizes. The experiments were conducted for the above scenarios for data acquisition and real-time testing, and the mAP of target detection of different models was counted, and the results are shown in Table 4. From Table 4, it can be seen that the model of this paper has the best detection effect on different targets in real scenarios, in which the detection mAP value for vehicles is as high as 89.32%, and the mAP value for all categories reaches 80.24%.

Table 4: Maps of Different Models(%)

Model	mAP	Car	Car plate	Pedestrian	Bus	Bicycle	Motorcycle	Tricycle
Faster R-CNN	69.32	75.34	58.44	67.35	75.23	62.12	70.35	73.23
Cascade R-CNN	68.42	73.23	57.23	68.34	70.34	61.94	69.43	60.42
Sparse R-CNN	70.23	79.34	60.32	74.34	73.52	64.28	72.48	53.55
RetinaNet	71.44	80.21	61.34	73.42	68.32	60.24	71.22	61.29
YOLOv5s	69.75	78.23	57.55	68.42	63.21	62.18	72.49	62.48
YOLOv5m	73.53	81.22	62.32	70.32	65.23	64.15	36.34	70.31
YOLOv3	74.86	82.43	65.23	72.43	70.53	62.48	62.93	68.32
YOLOv4	75.93	83.88	64.17	75.23	71.84	64.24	54.21	69.11
YOLOv3-Tiny	40.54	50.12	32.18	39.21	69.34	52.22	67.32	49.23
YOLOv4-Tiny	48.12	58.23	39.29	48.43	67.42	50.19	64.15	50.97
YOLOv7-Tiny	59.33	67.38	48.38	59.28	68.23	69.43	64.27	62.44
YOLO-MAXNet	70.42	78.32	61.15	69.23	72.43	70.11	63.55	70.21
YOLOv8s	71.43	79.22	62.48	70.35	79.89	62.48	64.18	72.18
The proposed	80.24	89.32	70.21	80.22	88.31	74.12	79.25	78.66

IV. Conclusion

In this paper, we design a real-time target detection algorithm for complex traffic scenes based on Yolov5s algorithm, and explore its performance level through dataset experiments and empirical analysis.

On the TT100K dataset, the accuracy of the model in this paper outperforms all comparison models, showing a significant improvement in performance compared to the YOLOv5s baseline model, with mAP@0.5 increasing by 22.92%, reaching an accuracy of 86.17%. Furthermore, this model only adds an additional 2.07M parameters. On the BDD100K dataset, the model achieves AP values of 74.58% and 69.21% for traffic signs and traffic lights, respectively, demonstrating optimal results in small object detection compared to other detection models. At the same time, this model maintains a lightweight parameter count and provides real-time detection speed, with mAP and parameter counts of 80.08% and 10.31M, respectively. Ablation experiments show that after using MHSARM, the model's AP@0.5 improved from a baseline of 63.25% to 75.23%, making the improved model perform best in AP@0.5 accuracy. The experimental results fully demonstrate that the architectural design of this model can enhance performance.

Experiments are conducted for multiple scenarios for data collection and real-time testing, the model in this paper has the best detection effect on different targets in real scenarios, in which the detection mAP value of vehicles is as high as 89.32%, and the mAP value of all categories reaches 80.24%.

References

[1] Vorozheikin, I., Marusin, A., Brylev, I., & Vinogradova, V. (2019, September). Digital technologies and complexes for provision of vehicular traffic safety. In International Conference on Digital Technologies in Logistics and Infrastructure (ICDTLI 2019) (pp. 380-384). Atlantis Press.

[2] Jain, N. K., Saini, R. K., & Mittal, P. (2019). A review on traffic monitoring system techniques. Soft computing: Theories and applications: Proceedings of SoCTA 2017, 569-577.

[3] Liu, T., Zhang, L., & Ding, X. (2025). Many-Objective Evolutionary Algorithms for Optimization of Vehicle-Road Cooperation Systems Based on Intelligent Wireless Sensor Networks. IEEE Internet of Things Journal.

[4] Gao, B., Liu, J., Zou, H., Chen, J., He, L., & Li, K. (2024). Vehicle-road-cloud collaborative perception framework and key technologies: A review. IEEE Transactions on Intelligent Transportation Systems.

[5] Gao, C., Wang, J., Lu, X., & Chen, X. (2022). Urban Traffic Congestion State Recognition Supporting Algorithm Research on Vehicle Wireless Positioning in Vehicle–Road Cooperative Environment. Applied Sciences, 12(2), 770.

[6] Yu, G., Li, H., Wang, Y., Chen, P., & Zhou, B. (2022). A review on cooperative perception and control supported infrastructure-vehicle system. Green Energy and Intelligent Transportation, 1(3), 100023.

[7] Park, M. W., In Kim, J., Lee, Y. J., Park, J., & Suh, W. (2017). Vision-based surveillance system for monitoring traffic conditions. Multimedia Tools and Applications, 76, 25343-25367.

[8] Datondji, S. R. E., Dupuis, Y., Subirats, P., & Vasseur, P. (2016). A survey of vision-based traffic monitoring of road intersections. IEEE transactions on intelligent transportation systems, 17(10), 2681-2698.

[9] Bai, J., Li, S., Zhang, H., Huang, L., & Wang, P. (2021). Robust target detection and tracking algorithm based on roadside radar and camera. Sensors, 21(4), 1116.

[10] Zheng, H., Liu, J., & Ren, X. (2022). Dim target detection method based on deep learning in complex traffic environment. Journal of Grid Computing, 20(1), 8.

[11] Bernas, M., Płaczek, B., Korski, W., Loska, P., Smyła, J., & Szymała, P. (2018). A survey and comparison of low-cost sensing technologies for road traffic monitoring. Sensors, 18(10), 3243.

[12] Pan, Q., & Zhang, H. (2020). Key algorithms of video target detection and recognition in intelligent transportation systems. International Journal of Pattern Recognition and Artificial Intelligence, 34(09), 2055016.

[13] Yang, B., & Zhang, H. (2022). A CFAR algorithm based on Monte Carlo method for millimeter-wave radar road traffic target detection. Remote Sensing, 14(8), 1779.

[14] Khan, H., & Thakur, J. S. (2024). Smart traffic control: machine learning for dynamic road traffic management in urban environments. Multimedia Tools and Applications, 1-25.

[15] Mistry, S., & Degadwala, S. (2023). A Comprehensive Review on Object Detectors for Urban Mobility on Smart Traffic Management. International Journal of Scientific Research in Computer Science, Engineering and Information Technology.

[16] Mahaur, B., & Mishra, K. K. (2023). Small-object detection based on YOLOv5 in autonomous driving systems. Pattern Recognition Letters, 168, 115-122.