IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 18 January 2023, accepted 28 February 2023, date of publication 3 March 2023, date of current version 8 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3252021

RESEARCH ARTICLE

MCS-YOLO: A Multiscale Object Detection Method for Autonomous Driving Road Environment Recognition

YINING CAO[1], **CHAO LI**[1], **YAKUN PENG**[1], **AND HUIYING RU**[2]

[1]College of Information Engineering, Hebei University of Architecture, Zhangjiakou 075000, China
[2]College of Science, Hebei University of Architecture, Zhangjiakou 075000, China

Corresponding author: Chao Li (1104876024@qq.com)

This work was supported in part by the 2021 'the 14th Five Year Plan' Project of Hebei Higher Education Association under Grant GJXHZ2021-22 and Grant GJXH2021-105, in part by the 2021 Research on the Development of Military Civilian Integration in Hebei Province under Grant HB21JMRH015, and in part by the Science and Technology Research Project of Colleges and Universities in Hebei Province under Grant QN2022097.

ABSTRACT Object detection and recognition of road scenes are crucial tasks of the autonomous driving environmental perception system. The low inference speed and accuracy in object detection models hinder the development of autonomous driving technology. Searching for improvement of detection accuracy and speed is still a challenging task. For solving these problems, we proposed an MCS-YOLO algorithm. Firstly, a coordinate attention module is inserted into the backbone to aggregate the feature map's spatial coordinate and cross-channel information. Then, we designed a multiscale small object detection structure to improve the recognition sensitivity of dense small object. Finally, we applied the Swin Transformer structure to the CNN to enable the network to focus on contextual spatial information. Conducting ablation study on the autonomous driving dataset BDD100K, MCS-YOLO algorithm achieves a mean average precision of 53.6% and a recall rate of 48.3%, which are 4.3% and 3.9% better than the YOLOv5s algorithm respectively. In addition, it can achieve real-time detection speed of 55 frames per second in a real scene. The results show that the MCS-YOLO algorithm is effective and superior in the task of automatic driving object detection.

INDEX TERMS Coordinate attention mechanisms, autonomous driving, road environmental object detection, swin transformer, YOLOv5.

I. INTRODUCTION

In the 21 century, cars have become an indispensable means of transportation and transport for people. The number of new vehicle registrations and newly licensed drivers worldwide is proliferating. The rapid increase in the number of motor vehicles has also brought about problems such as traffic accidents, traffic congestion, and environmental congestion. Autonomous driving technology is of great importance in resolving safety issues and decision-making in route planning during the driving process of motor vehicles [1], [2]. The primary task of the environmental perception system is to identify object information in the road environment precisely and quickly. After that, it will feed the identified information to the decision system to make the best route driving decision [3].

In the early development of autonomous driving technology, expensive single-sensor or multi-sensor fusion was employed to capture environmental object information. A technician sets the vehicle parameters subjectively and adjusts the parameters manually after several repeated simulations and field trials. At this stage, environmental perception technology requires significant human involvement and it is difficult to extend to new application scenarios [4]. With the rapid development of deep learning, sensing, and hardware technologies, Computer Vision (CV) and Natural Language Processing (NLP) fields have blossomed. Compared to sensor fusion methods, Deep Learning (DL) based object

The associate editor coordinating the review of this manuscript and approving it for publication was Abdullah Iliyasu.

detection algorithms can cost fewer computer resources and obtain higher detection accuracy, making it possible to satisfy real-time, accurate, and efficient environmental sensing systems.

Girshick et al. proposed the Regions with Convolutional Neural Networks Features (R-CNN) [5] model, which improved the recognition efficiency to a great extent. R-CNN converts the traditional object detection problem into a feature acquisition problem for regions and a classification problem for proposals. He et al. devised a Spatial Pyramid Pooling (SPP) [6] method to solve the problem of missing information in R-CNN models due to normalization. Fast R-CNN [7] reduces network training and testing time significantly. Faster R-CNN [8] uses Region Proposal Network (RPN) to extract bounding boxes, further improving the algorithm training and computing speed. Mask R-CNN [9] can perform detection and segmentation tasks with high quality.

The You Only Look Once (YOLO) series of algorithms [10], [11], [12], [13], [14], [15], [16] and the Single Shot MultiBox Detector (SSD) series of algorithms [17], [18], [19] adopt regression methods for object classification and bounding box prediction. The YOLO algorithm takes the entire image as input and regresses the position and class of the bounding box directly in the output layer. The YOLO and SSD algorithms are widely used in industry for their faster real-time detection than the R-CNN algorithm. Liu et al. used the Transformer as the backbone of a convolutional neural network for dense vision tasks. The success of the Swin Transformer [20], [21] demonstrates the powerful potential of the transformer for classification, detection and segmentation tasks. ConvNext [22] uses the same optimisation strategy as Swin Transformer to train the convolutional neural network. With the same FLOPs, ConvNext has faster inference and higher accuracy than Swin Transformer.

Chen et al. [23] proposed a DW-YOLO algorithm that improves vehicle object detection performance by increasing the depth and width of the network. Zhou et al. [24] proposed a lightweight MobileYOLO algorithm that reduces the number of parameters and improves detection speed. Wang et al. [25] applied MobileNet to a YOLOv4 network for driving scenarios and achieved a detection speed of 35 FPS. Tian et al. [26] proposed a SA-YOLOv3 detector that strikes a better compromise between detection speed and accuracy. Gupta et al. [27] applied both detection and segmentation to the task of road environment object detection to enhance the intelligent adaptive behaviour of self-driving cars. Wang et al. [28] propose an autonomous driving detection network for foggy weather that improves the accuracy of object detection in foggy weather scenarios as well as the speed of detection. Li et al. [29] designed a Res-YOLO network model that significantly reduced the missed-detection rate and improved the detection accuracy of vehicle object detection.

Object detection algorithms are constantly being improved and enhanced, demonstrating increasingly powerful performance. The success of state-of-the-art object detectors proves that techniques such as backbone network design and efficient feature fusion are essential to improve object detection performance. However, currently, problems such as low accuracy and inference speed hinder the development of autonomous driving technology in autonomous driving environment perception tasks. Achieving a compromise between inspection accuracy and speed is a challenging task. We propose a multiscale object detection algorithm for autonomous driving road environment object recognition. Our proposed algorithm achieves a bidirectional improvement in detection speed and accuracy for autonomous driving detection tasks.

The significant contributions of this paper are summarized as follows:

(1) Proposed a MCS-YOLO algorithm applied to the task of automatic driving object detection. We conducted ablation experiments and comparative trials of the MCS-YOLO algorithm on the BDD100K dataset. Compared to existing algorithms, the MCS-YOLO algorithm offers significant improvements in detection accuracy and speed.

(2) Designed a structure suitable for dense small object detection tasks. By using the new structure designed, the performance of the network in detecting small objects is effectively improved.

(3) Conducted several experiments to verify the effectiveness of the attention mechanism. The experimental results show that the coordinate attention mechanism performs best in the autonomous driving object detection task.

(4) Combining the Swin Transformer structure with CNNs enabled the network to have local relevance as well as global modelling capabilities.

The rest of this paper is organized as follows: Section II introduces attention mechanisms, multiscale feature fusion and real-time object detectors. In Section III, the MCS-YOLO algorithm is described in detail. Section IV, experimental datasets, parameter settings, and evaluation metrics are presented. Section V shows the results of ablation experiments and comparison experiments. The proposed algorithm is summarized, and future work is looked forward to in Section VI.

II. RELATED WORK

A. ATTENTION MECHANISM

Attention mechanisms have been proven to be an effective way to improve network performance. SENet [30] acquires the importance of feature channels by means of autonomous learning in order to establish dependencies between channels. Convolutional Block Attention Module (CBAM) [31] obtains the importance of the feature channels as well as the feature space via a similar approach. Global Attention Mechanism (GAM) [32] improves the performance of deep neural networks by reducing information diffusion and amplifying global interactions. The Acmix attention mechanism [33] integrates self-attention with convolution to further improve performance. In this paper, we have verified through several experimental comparisons that the Coordinate Attention (CA) mechanism [34] outperforms other attention

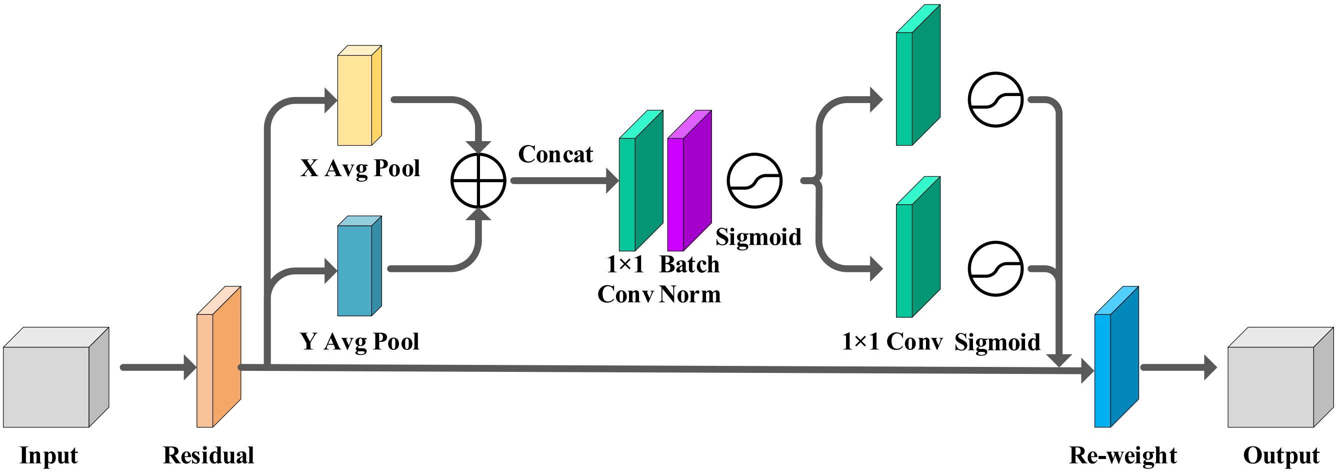


FIGURE 1. Coordinate Attention mechanism structure.

mechanisms for self-driving object detection tasks. And it does not increase the computational overhead of the network.

B. MULTISCALE FEATURE FUSION

The fusion of features between different scales is an important means to improve detection performance. Early object detectors used features extracted directly from the backbone. In contrast, Feature Pyramid Network (FPN) [35] combines multiscale features in a top-down approach. Path Aggregation Network (PANet) [36] adds a bottom-up vertical connection path to the FPN. NAS-FPN [37] uses a neural architecture search to find the best FPN structure. The Recursive Feature Pyramid (RFP) [38] adds feedback connection paths to enrich the feature representation. Although, NAS-FPN and RFP obtain an increase in performance, they have a large computational overhead. In this paper, we propose a multiscale feature fusion structure. In the structure, a detection layer suitable for small object detection tasks is added. And adding a horizontal spanning link to fuse multi-scale semantic information. Our proposed structure can effectively improve the network detection performance without increasing the computational effort. Experiments demonstrate that the structure we designed can be better applied to the task of self-driving target detection.

C. REAL-TIME OBJECT DETECTORS

Currently, object detectors are widely used in various computer vision tasks, such as classification, detection and segmentation. Efficient object detection classifiers are mainly based on CNN and Transformer. CNNs capture local features in a hierarchical manner for better feature maps, but have limitations in capturing global feature representations. Dosovitskiy et al. [39] used the transformer structure for dense visual tasks by constructing a hierarchical feature map, which has achieved exciting results. The success of the transformer in vision tasks is largely due to the global modelling capability over long distances. In this paper, taking the speed and accuracy of the model into consideration, we use YOLOv5s network as the basis. We combine the CNN network with the Transformer structure so that the network has local relevance as well as capturing feature dependencies over long distance. The proposed algorithm is applied to an autonomous driving object detection task to improve detection accuracy while maintaining a high real-time detection speed.

III. MCS-YOLO

A. COORDINATE ATTENTION MECHANISM

The dense small objects present in the road environment occupy less pixel information and are vulnerable to background factors. The YOLOv5 network tends to lose small objects information when convolutional sampling of them is performed. The coordinate attention mechanism is introduced in the MCS-YOLO algorithm to focus the network on crucial content and location information. For smaller objects, the location and spatial information can be extracted effectively to improve the accuracy of network detection.

The spatial and channel features are equally crucial for the generation of feature maps in the process of feature extraction by the network. Squeeze-and-Excitation (SE) block causes loss of spatial location information in global encoding operations. Bottleneck Attention Module (BAM) [40] and CBAM cannot obtain comprehensive range dependence information. Coordinate attention can solve the above problem well. Coordinate attention decomposes the global encoding from channel attention into one-dimensional parallel encodings along the horizontal and vertical directions. For an input feature map, the coordinate attention mechanism aggregates the position-aware features in each of the two directions. Each location-aware feature has cross-channel dependencies along the feature map in that direction. Coordinate attention enhances the network's perception of spatial location information and solves the problem of missing location information. The coordinate attention operation consists of two main processes: information embedding and attention generation.

As shown in Fig. 1., the coordinate attention mechanism decomposes the global pooling operation into pooling encoding operations along the horizontal and vertical directions. Global pooling is described as (1). For the input feature map X with dimension $C \times H \times W$, a pooling kernel of sizes $(H, 1)$ and $(1, W)$ is used to encode it. The feature map is then output with respect to the x and y axes and the location information is aggregated along the spatial direction of the features. The generated feature maps Z_c^h and Z_c^w as shown in (2) and (3).

$$Z_c^h = \frac{\sum_{i=1}^{H}\sum_{j=1}^{W} x_c(i,j)}{H \times W} \tag{1}$$

$$Z_c^h(h) = \frac{\sum_{0\leq i\leq w} x_c(h,i)}{W} \tag{2}$$

$$Z_c^w(w) = \frac{\sum_{0\leq j\leq H} x_c(j,w)}{H} \tag{3}$$

Meanwhile, coordinate attention gets the long-term dependence of spatial direction and keeps the position information in another spatial direction. Z_c^h and Z_c^w obtain the global sensory field of the input feature map with the exact location encoding information. Both are stitched together in the spatial dimension for the operation. The number of channels is then compressed by 1×1 convolution to obtain the attention feature map f, defined as (4).

$$f = \delta(F_1[Z_\text{c}^h, Z_c^w]) \tag{4}$$

where $[\cdot,\cdot]$ denotes a stitching operation along the spatial dimension. δ is a non-linear activation function. $f \in R^{C\times(H+W)/r}$ is a feature mapping of spatial information in the horizontal and vertical directions. Feature mapping encodes spatial information in the horizontal and vertical directions through batchnorm and non-linear operations.

Then, slice f into two different tensors, f^h and f^w, along the spatial dimension. And then transform the feature maps f^h and f^w by two 1×1 convolutions F_h and F_w to the same number of channels as the input feature map X. The results are obtained as follows:

$$g_c^h = \sigma(F_h(f^h)) \tag{5}$$

$$g_c^w = \sigma(F_w(f^w)) \tag{6}$$

g_c^h and g_c^w are subjected to a sigmoid activation function and then weighted with the original input information in both directions to obtain the result. The final output of the coordinate attention mechanism module can be expressed as shown in (7).

$$y_c(i,j) = x_c(i,j) \times g_c^h(i) \times g_c^w(j) \tag{7}$$

The coordinate attention mechanism embeds location information into channel attention, allowing lightweight networks to focus on a larger area. Meanwhile, the coordinate attention mechanism avoids incurring significant computational overhead. Coordinate attention drives the network to focus on small objects and objects containing fewer features present in the object detection task. Features and information that are more useful to the network for recognition are obtained.

B. MULTISCALE SMALL OBJECT DETECTION STRUCTURE

During the forward calculation stage of the YOLOv5s algorithm, the Conv module plays the role of down-sampling. Moreover, the C3 module focus on learning the residual features. After each standard convolution operation, the size of the output feature map will reduce by half. Thus, five feature maps are generated in the feature extraction process: $\{C_1, C_2, C_3, C_4, C_5\}$. The YOLOv5s algorithm combines FPN and PAN structures for feature fusion to identify objects of different scales. The FPN architecture merges the features obtained by up-sampling with $\{C_3, C_4, C_5\}$ through the Concat operation. At this time, the produced feature representations $\{P_3, P_4, P_5\}$ contain abstract high-resolution semantic information and the underlying positioning detail information. The PAN structure adds a bottom-up path based on the Feature Pyramid Networks. $\{N_i\}$ and $\{P_{i+1}\}$ are then fused by a Concat operation to obtain $\{N_{i+1}\}$.

Fig. 2. shows the YOLOv5s network feature extraction process and feature fusion process. It can be observed that the original YOLOv5s network has three different sizes of output. One of the 80 × 80 feature maps is used to detect small-size targets. For an image with an input size of 640 × 640, the receptive field size of one grid in the feature map is 8 × 8. It is difficult for the network to learn feature information for targets less than 8 pixels tall or wide in the original image. The trained model also has difficulty detecting such targets, leading to an excessive missed detection rate.

To improve the network's friendliness for smaller object detection, we have designed a small object detection structure, as shown in the blue background box section in Fig. 2. We continue to perform convolution operations, feature extraction, and up-sampling operation on the feature map $\{P_3\}$ to further expand the feature map. Then the output is fused with the feature map $\{C_2\}$ to generate $\{P_2\}$ with a size of 160 × 160. After feature extraction by the C3 module, $\{N_2\}$ is used as the output layer to detect small targets with a size over 4 × 4, which is equivalent to $\{P_2\}$. In addition, we add a shallow to deep spanning connection to the YOLOv5s network structure. As shown in the red arrow in Fig. 2., $\{C_2, C_3, C_4\}$ and $\{N_2, N_3, N_4\}$ are subjected to Concat operation. In this way, the detailed information of the feature map can be supplemented in space, and more accurate features can be extracted in the following sampling process, which is conducive to the detection of dense small targets.

C. SWIN TRANSFORMER LAYER

The transformer is not only powerful in its ability to model global contextual information, but large-scale pretraining also shows excellent transferability to downstream tasks. It has widely witnessed the success of the transformer in machine translation and natural language processing and provides new possibilities for visual feature learning.

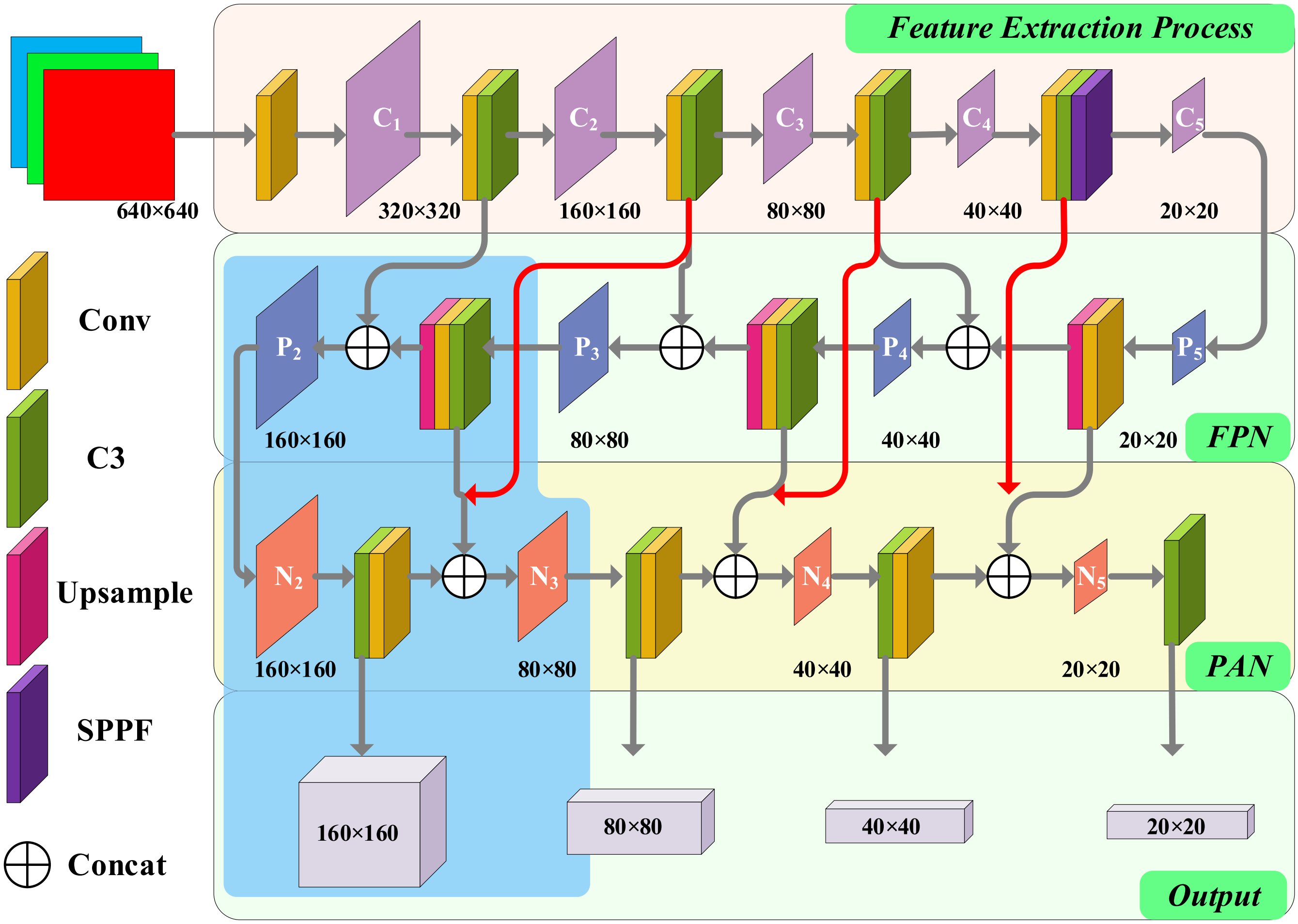


FIGURE 2. Multiscale small object detection structure.

The transformer constructs a global information interaction mechanism that helps to establish an adequate representation of features. However, there are two significant problems with applying the transformer to vision tasks. On the one hand, the expensive computational cost of the transformer, which uses sequences as input, dramatically limits its application to high-resolution input and intensive prediction tasks. On the other hand, unlike local inductive bias in convolution, transformer mines correlations from global relationships and requires training with large amounts of data to have excellent results.

The emergence of the swin tranformer opens up new possibilities for the application of transforms to visual tasks. The swin transformer has a small computational overhead. It processes images by constructing hierarchical structures, so that the transformer model can handle multi-scale intensive tasks. We applied the swin transformer structure to the YOLOv5s network structure, allowing the network to have global modelling capabilities while spending less computational resources. Fig. 3. demonstrates the swin transformer block structure. Swin transformer proposed the Multi-head Self Attention module for Windows (W-MSA). The image is divided into multiple windows. Swin transformer performs attention calculations on only the window pixel regions, reducing the computational complexity to a linear relationship. Crucially, the swin transformer operates using a Multi-head Self-Attention module for Shifted Windows (SW-MSA) for information interaction between non-overlapping windows.

The locality is a typical characteristic of CNN, an inductive bias based on the assumption that neighboring pixels have a significant correlation. CNN extracts features by sharing convolution kernels, dramatically reducing the number of parameters to improve the efficiency of network computing. On the other hand, the combination of convolution and pooling gives the network a certain translation invariance and translation equivalence. However, the limited perceptual field of CNN makes it challenging to capture global contextual information. In contrast, swin Transformer relies on more flexible self-attention information communication and shows excellent performance in extracting global semantic information and performance ceiling. Therefore, we combined CNN with the swin transformer, which helps the network to achieve

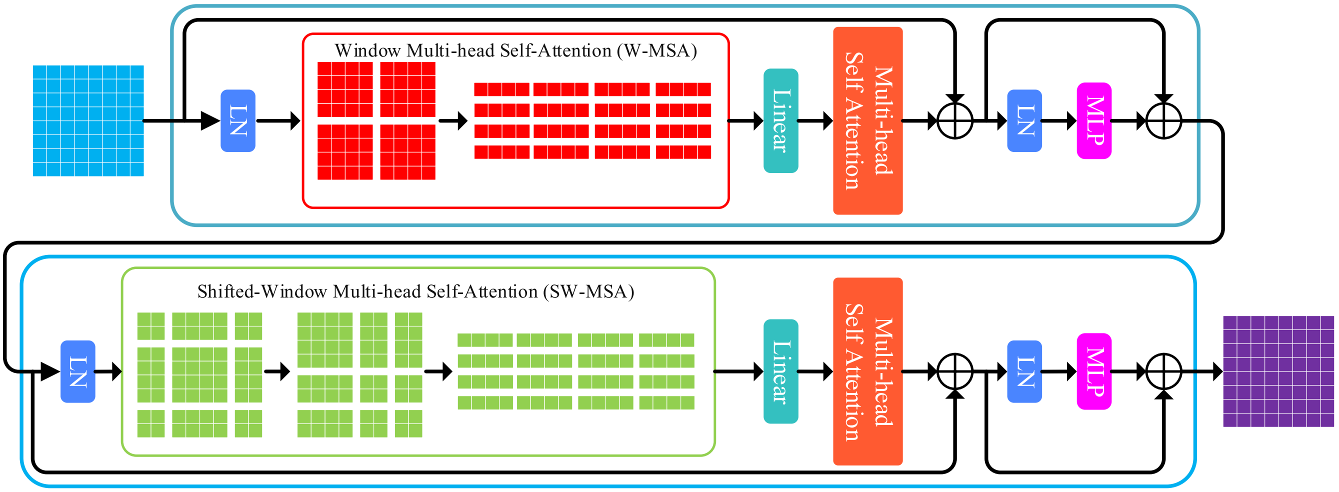

Window Multi-head Self-Attention (W-MSA)
LN
Linear
Multi-head Self Attention
LN
MLP
Shifted-Window Multi-head Self-Attention (SW-MSA)
LN
Linear
Multi-head Self Attention
LN
MLP

FIGURE 3. Swin transformer block.

interaction between local and global information and build an adequate feature representation.

D. MCS-YOLO STRUCTURE

Fig. 4. shows the overall structure of the MCS-YOLO network. The boxes in red represent the newly added modules. The YOLOv5s operates with convolution for downsampling and uses the C3 module for feature extraction. The features extracted by backbone are transmitted to the neck for feature fusion. The wealth of shallow location and channel information is critical to the diversity of deep feature fusion, which helps to boost the sensitivity of the model for road environment target recognition. Therefore, we embed the coordinate attention module before the SPPF layer in the YOLOv5s backbone to guide the allocation of different weights. Coordinate attention can fully use the channel and spatial information to promote feature perception in the channel. The YOLOv5s network with a coordinate attention module can effectively focus on the relevant feature information of small targets in the road environment, enabling the extraction of weak and minor features of targets.

In the actual scenario of autonomous driving environmental perception, there are more small and dense objects on peak travel periods or crowded streets. The accuracy of detecting dense small targets and the timeliness of making decisions play a decisive role in driving safety. The YOLOv5s network output detection head is not suitable for detecting targets with a size of more than 4 × 4. We have designed a structure suitable for dense small object detection. We added a new detection head, Detect0, to the YOLOv5s network to detect small targets. It will lose the shallower location information during the deepening process of YOLOv5s feature extraction. Therefore, attaching location information to the output layer is particularly essential. Taking it into account, we added a spanning connection from the backbone to the neck of the YOLOv5s network, as shown in the blue connection line in Fig. 4.

The convolution operation in CNN is local. It is challenging to obtain global semantic interaction and context information directly. To take advantage of the transformer's global modeling capabilities and avoid excessive memory overhead. We add the swin transformer layer behind the C3 module and use the swin transformer layer as output. The output detection heads now contain diverse local information and rich contextual interaction messages, which are crucial for enhanced detection performance.

IV. EXPERIMENT

A. DATASET

In order to verify the effectiveness and authenticity of the MCS-YOLO algorithm in the autonomous driving environment perception process, we use the authoritative public dataset BDD100K [41] for the evaluation experiment. The BDD100K dataset was collected in real-life scenarios, labeled with ten categories of targets in different weather, driving scenarios, and times of the day. The BDD100K dataset is large and diverse, containing 100,000 images. Moreover, the background of the dataset contains six different weather conditions: sunny, cloudy, cloudy, rainy, snowy, and foggy. To better validate the model's performance, we removed the 20,000 images in the dataset that did not contain labels and re-partitioned the remaining image data at an 8:1:1 ratio. The number of training sets is 64800, the number of validation sets is 7200, and the number of test sets is 8000. In Fig. 5., the object center points are primarily distributed in the central part of the image, the overall distribution of the object is uniform, and the small targets in the dataset account for a large proportion. Fig. 6. shows a graph of an example dataset.

B. EXPERIMENT PARAMETERS AND ENVIRONMENT SETTINGS

We conducted ablation experiments and comparison tests on the BDD100K dataset to verify the effectiveness of the MCS-YOLO algorithm. The operating system used to

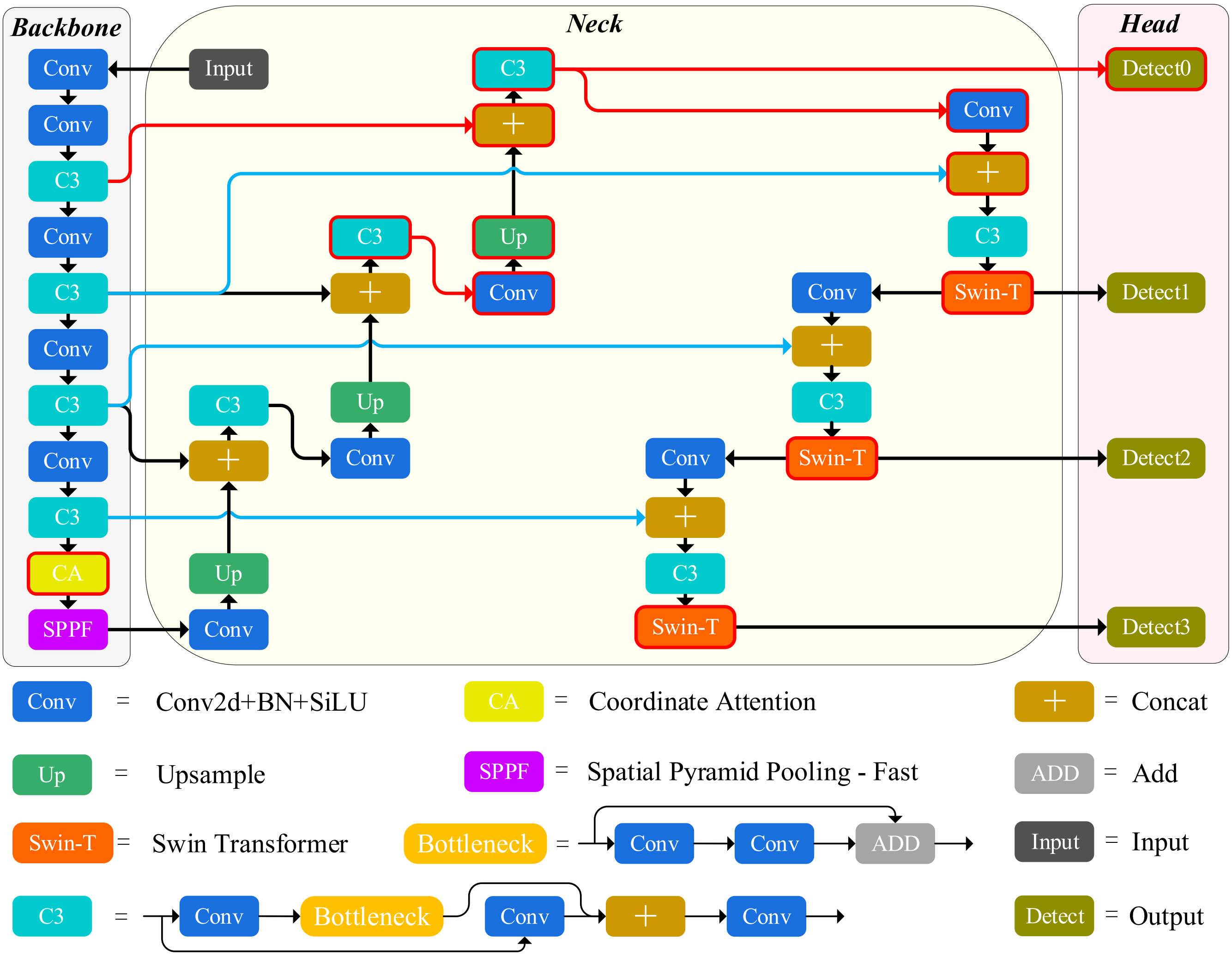
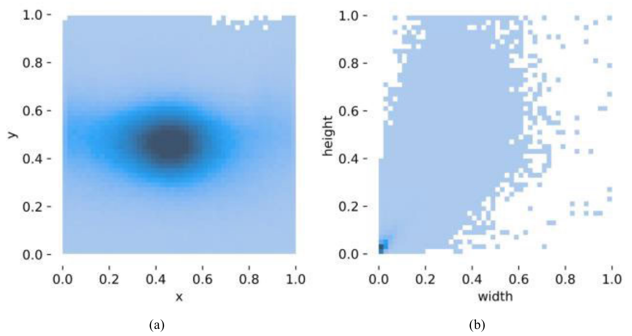


FIGURE 4. MCS-YOLO network architecture.



(a)

(b)

FIGURE 5. Location and size distribution of object center point.

perform the experiments is Ubuntu 18.04. The CPU model is Intel Xeon Platinum 8124M. The GPU model is GeForce RTX 3070 Ti. The programming language is Python 3.8.13. The acceleration environment is CUDA 11.4, and the deep learning framework is Pytorch 1.10.0.

Experiment parameters are set as shown in Table 1.

TABLE 1. Experiment parameter settings.

Parameters	Parameter settings
hyperparameters file	hyp. scratch-low. yaml
weights	yolov5s.pt
epochs	200
patience	30
batchsize	32
workers	12

V. RESULTS AND ANALYSIS

A. ANALYSIS OF ABLATION EXPERIMENTS

To verify the effectiveness and superiority of each improvement point in the proposed algorithm. We conducted ablation experiments on the test set divided by the BDD100K dataset. "$\surd$" represents the use of this modular method. Table 2 shows the experimental results.

The experiment results show that every improvement point in the MCS-YOLO algorithm improves performance for the

TABLE 2. Experimental results of MCS algorithm ablation on the test set of the BDD100K dataset.

Experimental programmes	Coordinate Attention	Multiscale Small Object Detection Structure	Swin Transformer Layer	mAP@.5(%)	mAP@.5:.95(%)	Precision	Recall	FPS
0				49.3%	26.5%	75.5%	44.4%	107
1	√			49.8%	26.7%	71.7%	45.1%	89
2		√		52.9%	28.2%	71.8%	47.9%	105
3			√	51.1%	27.2%	70.3%	46.8%	69
4	√	√		53.1%	28.1%	72.8%	47.1%	88
5	√	√	√	53.6%	28.6%	73.7%	48.3%	55

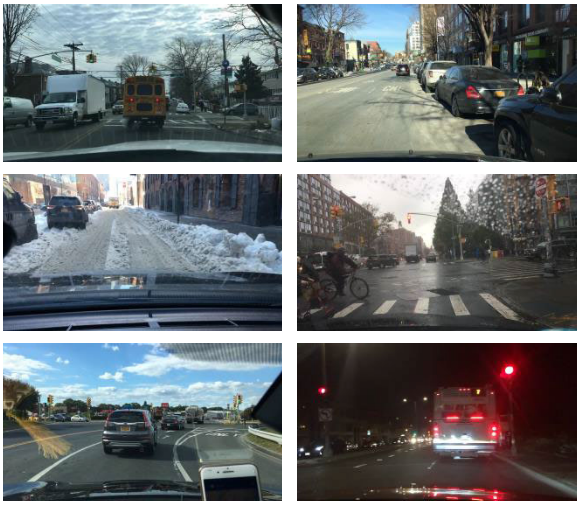


FIGURE 6. Example of a dataset image.

network compared to YOLOv5s. In scheme 2, we add a multiscale small object detection structure to the network. The test results show that mAP@.5 increases by 3.6%, mAP@.5:.95 increases by 1.7%, and recall increases by 3.5%. In scheme 3, we apply the swin transformer structure to the network. The test results show that mAP@.5 increases by 1.8%, mAP@.5:.95 increases by 0.7%, and recall increases by 2.4%.

Coordinate attention can focus on critical features of interest to the network, effectively improving the network's ability to aggregate features. The multiscale small object detection structure can easily capture small objects in the road environment and is highly sensitive to small targets. The swin transformer structure does self-attention operations within the window to better capture contextual semantic information for efficient extraction of global features.

B. ALGORITHM PERFORMANCE ANALYSIS

We compare the experimental results of scheme 0 and scheme 5 in Table 2, showing that the MCS-YOLO algorithm outperforms the YOLOv5s algorithm. Under the same experiment environment, all evaluation indicators of MCS-YOLO have been improved; mAP@.5 reached 53.6%, mAP@.5:.95 reached 28.6%, Precision reached 73.7%, and Recall reached 48.3%. The real-time detection speed reaches 55 frames per second. The MCS-YOLO algorithm model effectively improves detection accuracy, reduces the missed-detection rate of small targets, and meets real-time detection.

Fig. 7. shows the results of the single-class average accuracy comparison between the MCS-YOLO algorithm and the YOLOv5s algorithm. The single-class average precision of the MCS-YOLO algorithm is higher than that of YOLOv5s. Among them, the average precision of categories such as Traffic signs, traffic lights, persons, and cars has been dramatically improved because there are more dense small targets in the above categories. The MCS-YOLO algorithm can improve the accuracy of small targets and effectively reduce the missed detection rate. The results prove the effectiveness of the MCS-YOLO algorithm in object detection in autonomous driving road environments.

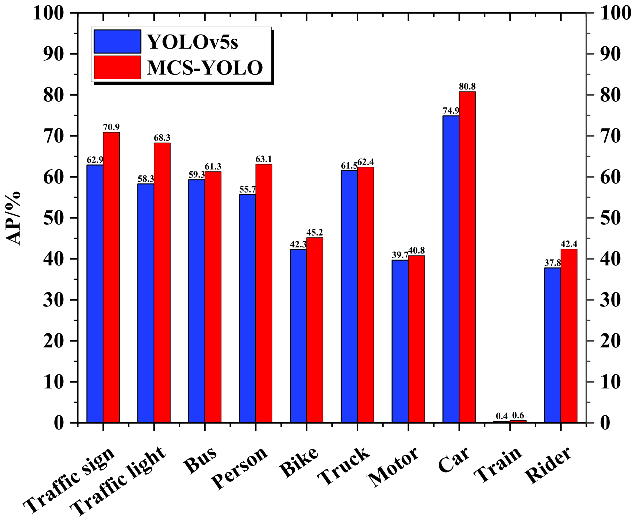


FIGURE 7. Single-class average precision comparison.

Fig. 8. shows the PR curve comparison between the MCS-YOLO algorithm and the YOLOv5s algorithm. The recall is the horizontal coordinate, and Precision is the vertical coordinate. The area enclosed by the PR curve and the

TABLE 3. Performance comparison of YOLOv5s and MCS-YOLO algorithms for road environment object detection of different sizes.

	IoU	Area	maxDets	YOLOv5s	MCS-YOLO
Average Precision (AP)	0.50:0.95	all	100	0.323	0.356
	0.50	all	100	0.602	0.664
	0.75	all	100	0.285	0.315
	0.50:0.95	small	100	0.147	0.187
	0.50:0.95	medium	100	0.421	0.447
	0.50:0.95	large	100	0.609	0.608
Average Recall (AR)	0.50:0.95	all	1	0.222	0.233
	0.50:0.95	all	10	0.432	0.469
	0.50:0.95	all	100	0.466	0.516
	0.50:0.95	small	100	0.314	0.389
	0.50:0.95	medium	100	0.593	0.613
	0.50:0.95	large	100	0.694	0.698

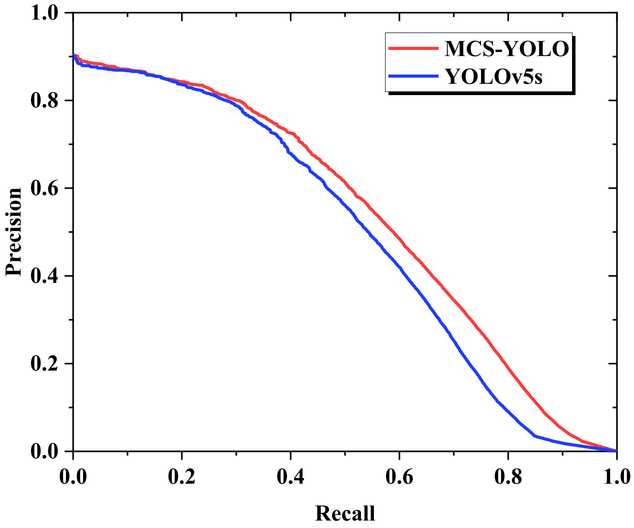


FIGURE 8. PR curve comparison.

horizontal and vertical axes is the average precision value. The PR curve of the MCS-YOLO algorithm encloses the PR curve of YOLOv5s, demonstrating the superior performance of the proposed algorithm. Fig. 9. shows the confusion matrix obtained for the YOLOv5s network model on the test set. Fig. 10. shows the confusion matrix obtained for the MCS-YOLO network model on the test set. The MCS-YOLO algorithm has a higher accuracy rate and a lower error and missed-detection rate. Fig. 11. shows the heat map comparison between the MCS-YOLO algorithm and the YOLOV5s algorithm. According to the heat map, it can be observed that the object receives more attention from the MCS-YOLO model than YOLOv5s. The results show that the MCS-YOLO algorithm can focus more on feature information, has greater sensitivity to detect objects, and performs better.

To further evaluate the superiority of the MCS-YOLO algorithm for the detection performance of small objects in the road environment. According to Microsoft's COCO benchmark evaluation metrics, we divided the road environment objects into small, medium and large objects. Among them, the area of small object is less than 32^2 pixels, the area of medium object is more than 32^2 pixels and less than 96^2 pixels, and the area of large object is more than 96^2 pixels.

TABLE 4. Experiment results comparing different algorithmic models on the BDD100K dataset test set.

Model	mAP (%)	FPS
Faster-RCNN [42]	43.1%	12.5
AD-Faster-RCNN [42]	50.8%	6
YOLOv3 [43]	40.1%	20.46
YOLOv3-tiny [43]	16.7%	143.92
H-YOLOv3 [43]	48.5%	35.14
YOLOv4 [44]	45.19%	3.89
Mobilenetv2-YOLOv4 [45]	49.97%	—
CDMY [45]	50.93%	—
YOLOv5s [46]	49.3%	60
Improved YOLOv5s [46]	51.2%	35
YOLOv7-tiny [16]	48.7%	278
MCS-YOLO	53.6%	55

In Table 3, for all detection objects, the MCS-YOLO detector has higher AP values and AR values in the case of different Intersection over Union (IoU). For different sizes of detected objects, the AP and AR values obtained by the MCS-YOLO algorithm are higher than those of the YOLOv5s algorithm for the same IoU. In particular, the AP and AR values for small objects increased more significantly. Specifically, the MCS-YOLO algorithm achieved an AP value of 18.7% for small object detection, an increase of 4 percentage points. The AR value of MCS-YOLO algorithm for small object detection reaches 38.9 %, which is increased by 7.5 percentage points. Experimental results illustrate the superior performance of the MCS-YOLO algorithm for small object detection in road environments.

C. COMPARATIVE EXPERIMENTAL ANALYSIS

Table 4 shows the experimental results of different algorithm models on the test set of the BDD100K dataset. The MCS-YOLO algorithm achieves a mean average precision of 53.6%, a 4.3% improvement over YOLOv5s, and a real-time detection speed of 55 frames per second. The results show that the MCS-YOLO algorithm outperforms the single-stage and two-stage algorithms, has higher detection accuracy, and can meet real-time detection.

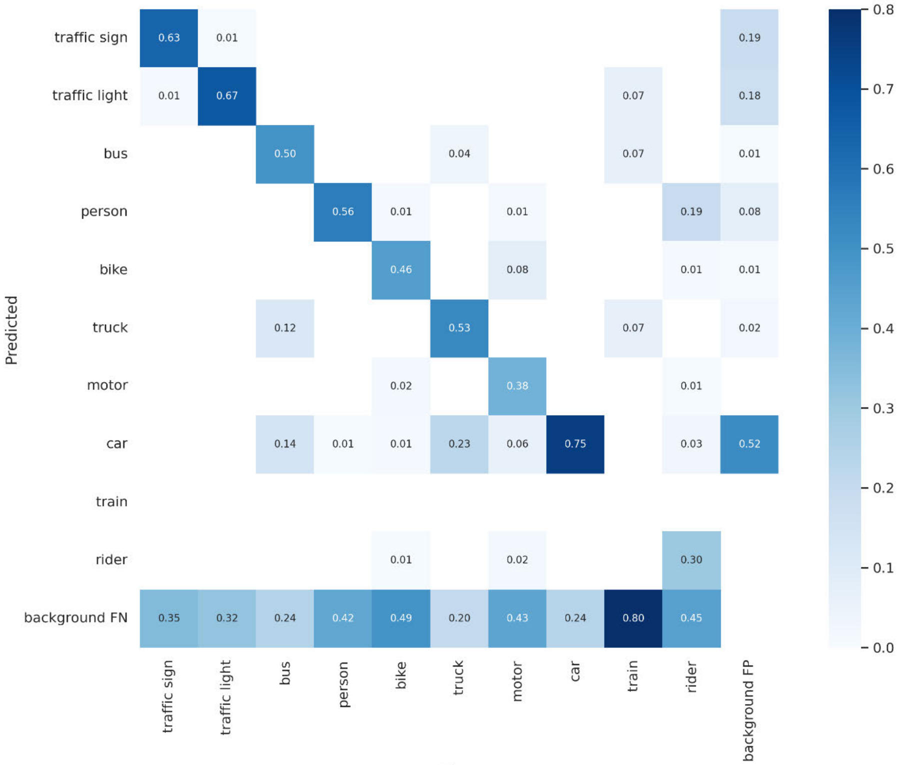


FIGURE 9. **Confusion matrix for YOLOv5s network model.**

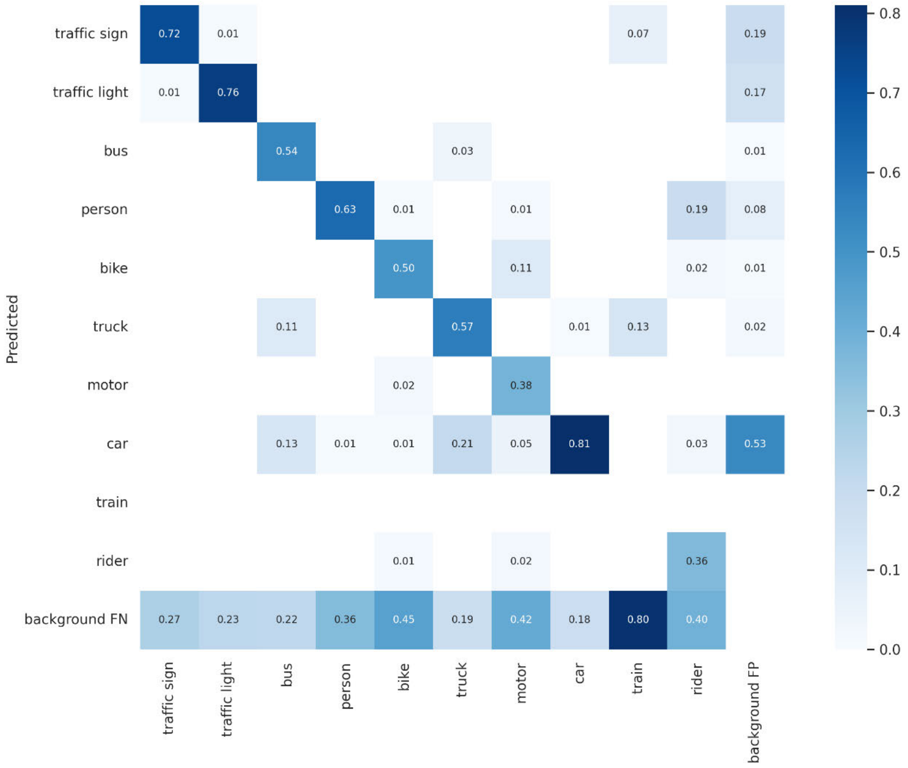


FIGURE 10. **Confusion matrix for MCS-YOLO network model.**

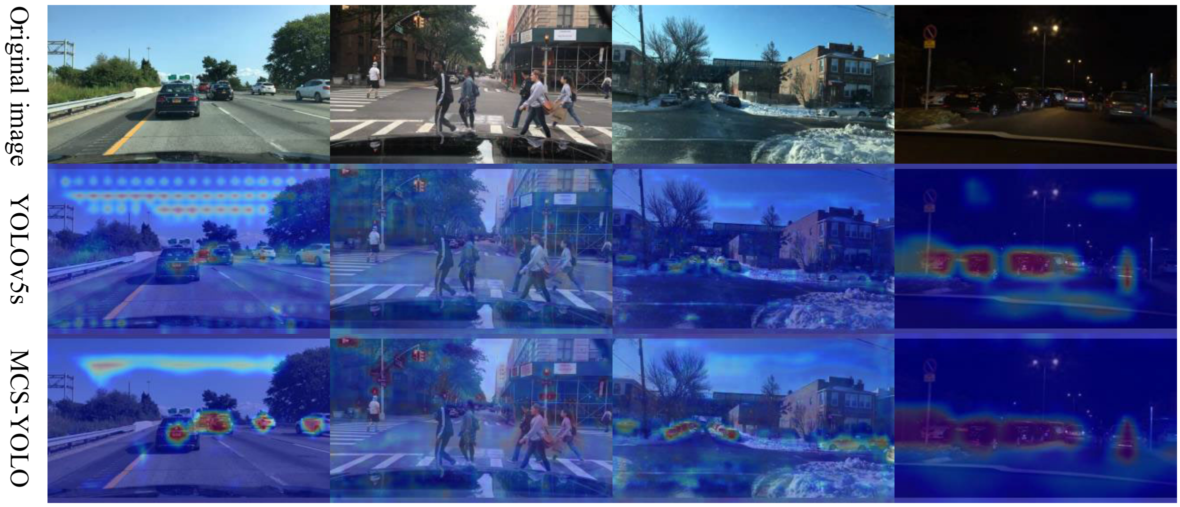


FIGURE 11. Heat map comparison.

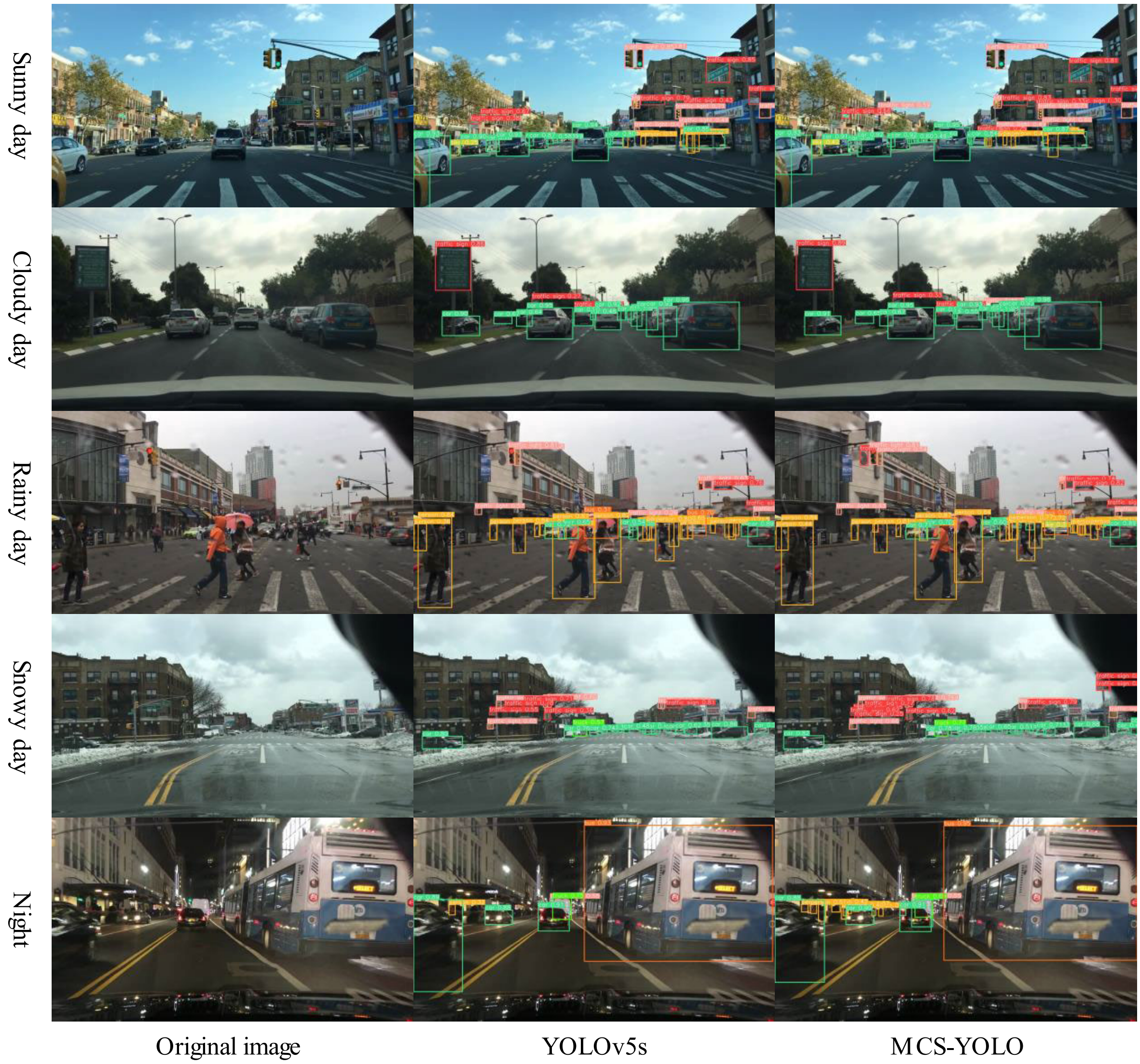


FIGURE 12. Comparison of actual test results.

Fig. 12. shows the actual detection results of the MCS-YOLO algorithm and YOLOv5s in different scenarios with different weather. It can be observed that the MCS-YOLO algorithm has better generalization and applicability in different weather and scenarios. It can detect more small targets in the road environment. In contrast, the YOLOv5s network is unable to detect smaller targets. Compared to the YOLOv5s algorithm, the MCS-YOLO algorithm has a higher confidence level for detecting the same object. The MCS-YOLO algorithm effectively reduces the missed detection and error rate, proving the effectiveness and superiority of the proposed algorithm. The MCS-YOLO algorithm is robust and can be applied to different autonomous driving scenarios with a higher accuracy rate.

VI. CONCLUSION

To address the problems of low inference speed and low object detection accuracy in autonomous driving road environments. We proposed a multiscale road object detection algorithm called MCS-YOLO. We introduced a coordinate attention mechanism in the backbone section to enhance the network's ability to aggregate features. We designed a multiscale small object detection structure that allows the network to perform multiscale detection of large, medium, small, and smaller targets. We apply the swin transformer structure to the neck part of the network, which can obtain accurate contextual feature information. The ablation experiments were conducted on the BDD100K dataset. The experiment results show that the MCS-YOLO algorithm achieves an average detection accuracy of 53.6% on the test set, an improvement of 4.3%. The MCS-YOLO model achieves 55 FPS in real scenarios, meeting the accuracy and real-time requirements for the detection of autonomous road environments. Compared with other mainstream algorithms, the MCS-YOLO algorithm is more suitable for application in autonomous driving environment perception tasks and has effectiveness and superiority. Multiple object tracking is an essential and challenging research in autonomous driving vision tasks. In future work, we will apply the MCS-YOLO algorithm to the autonomous driving Multiple Object Tracking (MOT) task to verify the performance of the proposed algorithm in this task.

REFERENCES

[1] J. Levinson, J. Askeland, J. Becker, J. Dolson, D. Held, S. Kammel, J. Z. Kolter, D. Langer, O. Pink, V. Pratt, M. Sokolsky, G. Stanek, D. Stavens, A. Teichman, M. Werling, and S. Thrun, ''Towards fully autonomous driving: Systems and algorithms,'' in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2011, pp. 163–168.

[2] J. Janai, F. Güney, A. Behl, and A. Geiger, ''Computer vision for autonomous vehicles: Problems, datasets and state of the art,'' *Found. Trends Comput. Graph. Vis.*, vol. 12, no. 1–3, pp. 1–308, 2020.

[3] Y. Wang, ''Overview on key technology of perceptual system on self-driving vehicles,'' *Auto Electr. Parts.*, vol. 2016, no. 12, pp. 12–16, Dec. 2016.

[4] M. Kuderer, S. Gulati, and W. Burgard, ''Learning driving styles for autonomous vehicles from demonstration,'' in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2015, pp. 2641–2646.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik, ''Rich feature hierarchies for accurate object detection and semantic segmentation,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[6] K. He, X. Zhang, J. Sun, and S. Ren, ''Spatial pyramid pooling in deep convolutional networks for visual recognition,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Jan. 2015.

[7] R. Girshick, ''Fast R-CNN,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[8] 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, Jun. 2017.

[9] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, ''Mask R-CNN,'' in *IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, ''You only look once: Unified, real-time object detection,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[11] J. Redmon and A. Farhadi, ''YOLO9000: Better, Faster, stronger,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 6517–6525.

[12] J. Redmon and A. Farhadi, ''YOLOv3: An incremental improvement,'' 2018, *arXiv:1804.02767*.

[13] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, ''YOLOv4: Optimal speed and accuracy of object detection,'' 2020, *arXiv:2004.10934*.

[14] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, ''YOLOX: Exceeding Yolo series in 2021,'' 2021, *arXiv:2107.08430*.

[15] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie, Y. Li, B. Zhang, Y. Liang, L. Zhou, X. Xu, X. Chu, X. Wei, and X. Wei, ''YOLOv6: A single-stage object detection framework for industrial applications,'' 2022, *arXiv:2209.02976*.

[16] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, ''YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors,'' 2022, *arXiv:2207.02696*.

[17] W. Liu, ''SSD: Single shot multibox detector,'' in *Proc. Eur. Conf. Comput. Vis.* Amsterdam, The Netherlands, 2016, pp. 21–37.

[18] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, ''DSSD : Deconvolutional single shot detector,'' 2017, *arXiv:1701.06659*.

[19] Z. Li and F. Zhou, ''FSSD: Feature fusion single shot multibox detector,'' 2017, *arXiv:1712.00960*.

[20] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, ''Swin transformer: Hierarchical vision transformer using shifted Windows,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[21] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, and B. Guo, ''Swin transformer v2: Scaling up capacity and resolution,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12009–12019.

[22] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, ''A ConvNet for the 2020s,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11976–11986.

[23] Y. Chen et al., ''DW-YOLO: An efficient object detector for drones and self-driving vehicles,'' *Arabian J. Sci. Eng.*, vol. 48, pp. 1–10, May 2022.

[24] Y. Zhou, S. Wen, D. Wang, J. Meng, J. Mu, and R. Irampaye, ''MobileYOLO: Real-time object detection algorithm in autonomous driving scenarios,'' *Sensors*, vol. 22, no. 9, p. 3349, Apr. 2022.

[25] H. Wang and W. Zang, ''Research on object detection method in driving scenario based on improved YOLOv4,'' in *Proc. IEEE 6th Inf. Technol. Mechatronics Eng. Conf. (ITOEC)*, Mar. 2022, pp. 1751–1754.

[26] D. Tian, ''SA-YOLOv3: An efficient and accurate object detector using self-attention mechanism for autonomous driving,'' *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 5, pp. 4099–4110, May 2022.

[27] A. Gupta, K. Illanko, and X. Fernando, ''Object detection for connected and autonomous vehicles using CNN with attention mechanism,'' in *Proc. IEEE 95th Veh. Technol. Conf., (VTC-Spring)*, Jun. 2022, pp. 1–6.

[28] H. Wang, Y. Xu, Y. He, Y. Cai, L. Chen, Y. Li, M. A. Sotelo, and Z. Li, ''YOLOv5-fog: A multiobjective visual detection algorithm for fog driving scenes based on improved YOLOv5,'' *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–12, 2022.

[29] Y. Li, J. Wang, J. Huang, and Y. Li, ''Research on deep learning automatic vehicle recognition algorithm based on RES-YOLO model,'' *Sensors*, vol. 22, no. 10, p. 3783, May 2022.

[30] J. Hu, L. Shen, and G. Sun, ''Squeeze-and-excitation networks,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[31] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–19.

[32] Y. Liu, Z. Shao, and N. Hoffmann, "Global attention mechanism: Retain information to enhance channel-spatial interactions," 2021, *arXiv:2112.05561*.

[33] X. Pan, C. Ge, R. Lu, S. Song, G. Chen, Z. Huang, and G. Huang, "On the integration of self-attention and convolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 815–825.

[34] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13708–13717.

[35] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.

[36] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8759–8768.

[37] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "NAS-FPN: Learning scalable feature pyramid architecture for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7036–7045.

[38] S. Qiao, L.-C. Chen, and A. Yuille, "DetectoRS: Detecting objects with recursive feature pyramid and switchable atrous convolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10213–10224.

[39] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[40] J. Park, S. Woo, J. Lee, and I. Kweon, "BAM: Bottleneck attention module," presented at the 29th Brit. Mach. Vis. Conf., Newcastle, U.K., Sep. 2018.

[41] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "BDD100K: A diverse driving dataset for heterogeneous multitask learning," 2018, *arXiv:1805.04687*.

[42] Y. Zhou, S. Wen, D. Wang, J. Mu, and I. Richard, "Object detection in autonomous driving scenarios based on an improved faster-RCNN," *Appl. Sci.*, vol. 11, no. 24, Dec. 2021, Art. no. 11630.

[43] L. Li and X. Li, "H-YoLov3: High performance object detection applied to assisted driving," in *Proc. Asia Conf. Algorithms, Comput. Mach. Learn. (CACML)*, Mar. 2022, pp. 462–467.

[44] F. Yang, X. Zhang, S. Zhang, C. Li, and H. Hu, "Design of real-time vehicle detection based on YOLOv4," in *Proc. Int. Conf. Control, Autom. Inf. Sci. (ICCAIS)*, Oct. 2021, pp. 824–829.

[45] Y. Li and F. Yu, "CDMY: A lightweight object detection model based on coordinate attention," in *Proc. IEEE 10th Joint Int. Inf. Technol. Artif. Intell. Conf. (ITAIC)*, Jun. 2022, pp. 1258–1263.

[46] Q. Luo, J. Wang, M. Gao, Z. He, Y. Yang, and H. Zhou, "Multiple mechanisms to strengthen the ability of YOLOv5s for real-time identification of vehicle type," *Electronics*, vol. 11, no. 16, p. 2586, Aug. 2022.

YINING CAO was born in Cangzhou, Hebei, in 1997. He received the B.S. degree from the Hebei University of Architecture, in 2019, where he is currently pursuing the M.S. degree with the College of Information Engineering. His research interests include computer vision and object detection.

CHAO LI was born in 1984. He received the B.S. degree in computer science and technology from Beijing Normal University, in 2007, and the M.S. degree in applied mathematics and the Ph.D. degree in management science from Liaoning Technical University, in 2011 and 2017, respectively. Since 2017, he has been an Associate Professor with Hebei University of Architecture, Zhangjiakou, China. He has published over 15 papers in refereed journals and conferences in his research areas and holds ten software copyrights. His research interests include optimization and artificial intelligence and recent research on fixed point algorithms and deep learning. He is a member of the China Computer Federation (CCF).

YAKUN PENG was born in Zhangjiakou, Hebei, in 1995. She received the B.S. degree from Hebei University of Science and Technology, in 2019. She is currently pursuing the M.S. degree with the College of Information Engineering, Hebei University of Architecture. Her research interests include image processing, computer vision, and object detection.

HUIYING RU was born in 1987. She received the B.S. degree in mathematics and applied mathematics from Beijing Normal University, in 2007, and the M.S. degree in applied mathematics from Liaoning Technical University, in 2016. Since 2017, she has been a Lecturer with Hebei University of Architecture, Zhangjiakou, China. Her research interests include optimization and artificial intelligence and recent research on fixed point algorithms and deep learning.