Journal of Computers Vol. 33 No. 1, February 2022, pp. 1-11
doi: 10.53106/199115992022023301001

MASFF: Multiscale Adaptive Spatial Feature Fusion Method for Vehicle Recognition

Jiefei Zhang[1*]

[1] School of Automobiles, Henan College of Transportation, Zhengzhou 450000, China
xdwangxd@163.com

Received 27 September 2021; Revised 15 October 2021; Accepted 27 October 2021

Abstract. Traditional vehicle recognition methods have the disadvantages such as low efficiency and time-consuming due to the complex background and overlapping situation. In this paper, we propose a multiscale adptive spatial feature fusion (ASFF) method for vehicle recognition. First, it calculates the difference hash values of images. Then the hash value is used to judge the similarity between the current frame and the previous frame. When the similarity is less than the threshold value, it is input to ResNet18 model for detection. Using ResNet18 as the base network can reduce network parameters. Then, aiming at the problem that the detection effect of ASFF for vehicle recognition is not ideal, the offset loss and width-height loss are replaced by the intersection ratio loss. Meanwhile, multi-scale adaptive spatial feature fusion method is adopted to fuse the multi-level features of the network. The experimental results show that the average accuracy with proposed methed is increased by 2.1%. For BDD100K and Pascal VOC datasets, the average accuracy of predicted borders is increased by 5.5%, when the IoU is greater than 0.5. With the GTX1060Ti, the recognition speed can reach 149 frames per second. The multiscale ASFF in this paper can significantly improve the vehicle recognition accuracy.

Keywords: vehicle recognition, multiscale adptive spatial feature fusion, ResNet18, hash value

1 Introduction

With the continuous advancement of urbanization, the number of urban population and the number of urban motor vehicles increase rapidly, which brings great pressure to urban public security. In order to effectively crack down on illegal behaviors and ensure city safety, intelligent monitoring systems are built in key sections and places in the city to effectively supervise pedestrians, motor vehicles and non-motor vehicles. Vehicle recognition has become one of the important functions in intelligent monitoring system [1,2]. Vehicle recognition is to detect the position of the vehicle object in the image and identify the object type. In practice, due to factors such as background clutter, illumination conditions, vehicle deformation, partial occlusion and motion blur, the recognition accuracy of vehicle type is not high. In order to improve the accuracy, domestic and foreign scholars have carried out a lot of researches.

Traditional vehicle recognition methods mainly use image edge feature, color feature, scale invariant feature transform (SIFT) [3] and directional gradient histogram (HOG) feature [4] to recognize vehicle objects. Zhang et al. [5] used color feature transformation of images and Bayesian classifier to mark the location of vehicles for classification and recognition. Ambardekar et al. [6] classified vehicle images through image edge features and principal component analysis. These two methods have short computation time and can quickly extract contour features of images. However, their performance degrades sharply in the case of large vehicle scale changes, partial occlusion, illumination changes, etc., and they cannot correctly identify vehicles. Briz-Redón et al. [7] combined Sobel feature and SIFT feature to identify vehicles, but the feature dimension was high and the calculation was large. Arrospide et al. [8] used HOG to extract vehicle features, but the calculation cost was high and the feature expression ability was general. In view of the low performance of traditional detection and recognition methods, deep learning has gradually become the mainstream algorithm for vehicle detection and recognition. Yao et al. [9] conducted prior measurement of images to obtain vehicle positioning, and then used convolutional neural network (CNN) to classify and recognize vehicles. Huo et al. [10] proposed a region-based multi-task CNN model to identify vehicle types, orientation and lighting conditions. Traditional CNN algorithm requires preset vehicle location, which is not conducive to real-time vehicle recognition. Due to the significant breakthrough in object recognition made by regional convolutional neural network (RCNN) [11,12] series algorithms, researchers began to apply these algorithms to vehicle detection and recognition. Tang et al. [13] proposed the super-region candidate network (HRPN), which could improve the recognition effect of Faster-RCNN [14] on small object vehicles. In order to solve the problem of vehicle occlusion and deformation, Wang et al. [15] introduced

[*] Corresponding Author

adversation learning into vehicle detection and recognition, which could generate difficult samples (samples with serious occlusion and deformation of vehicles) and improve the recognition effect of Fast-RCNN. Hu et al. [16] designed SINet in view of the fact that the ROI pool would destroy the structure of small objects, which improved the recognition effect and speed of vehicle objects with different sizes. RCNN series algorithms belong to two-stage methods, which have high accuracy but slow speed, and cannot realize real-time object recognition in edge devices. In order to improve the speed of vehicle detection and recognition, one-stage method is introduced. As shown in references [17-19], YOLO series algorithms are applied to vehicle detection and recognition. Cao et al. [20] improved SSD network by designing cascade module and element addition module for feature fusion, which improved the accuracy of vehicle identification, but reduced the detection speed. This kind of algorithm based on YOLO or SSD requires a set of candidate boxes with specific aspect ratio in advance, so it is also called candidate box based algorithm. In this algorithm, the design of candidate box is difficult to take into account the number of minimal or maximal objects, so in practical application, it is easy to cause the omission of such objects. In view of the above problems, the one-stage object recognition method based on no candidate box has become the current research hotspot. For example, Chang et al. [21] combined HRNet [22] and CenterNet [23] for vehicle detection and recognition.

Aiming at the shortcomings of existing object detection algorithms without candidate boxes in vehicle recognition applications such as excessive negative samples, this paper proposes a multiscale adptive spatial feature fusion (MASFF) method for vehicle recognition. The validity and superiority of this new method are verified by vehicle model recognition dataset, BDD100K dataset and Pascal VOC public dataset.

2 The Proposed MASFF

First of all, each frame of input video is processed by different hash algorithm to get the hash value. Then the Hamming Distance is calculated and compared with the threshold value to judge the similarity between this frame image and the previous frame image. Finally, the algorithm is selected according to the size of similarity so as to improve the average calculation speed of the whole algorithm. When the similarity is less than the threshold value, it is input to ResNet18 model for detection.

2.1 Different Hash Algorithm (DHA)

The steps of differential hash algorithm (DAH) [24] are as follows:

generate differential hash images D_i and D_j by neighborhood block N_i and N_j.

The hamming distance $D(i,j)$ is obtained by XOR operation of D_i and D_j.

$$D(i,j) = X_{XOR}(D_i, D_j) \ . \tag{1}$$

The hamming distance weight of gradient image is used to adjust the Euclidean distance weight in the experiment, and the Sobel operator of image gradient is calculated.

2.2 MASFF

Real-time and accuracy are the key factors of vehicle identification. Because ASFF is based on one stage object recognition model without candidate box, the inference speed is fast. In this paper, ASFF network is modified to improve the accuracy of the network in vehicle recognition. As shown in Fig. 1, the improved ASFF primarily consists of a backbone network, bottleneck modules, and an output network. In the backbone network, ResNet18 model has few parameters, so this paper chooses it as the basic network. In the bottleneck module, two feature fusion methods (single-scale adaptive spatial feature fusion (SSASFF) and multiscale adaptive feature fusion (MAFF)) are proposed to fuse multiscale features of the network to improve the identification performance of the network. In the output network part, the network predicts the classification category, the length/width of the frame and the bias of the center point of the identified object respectively. And it directly regresses all the information of the predicted object without using the model prior box.

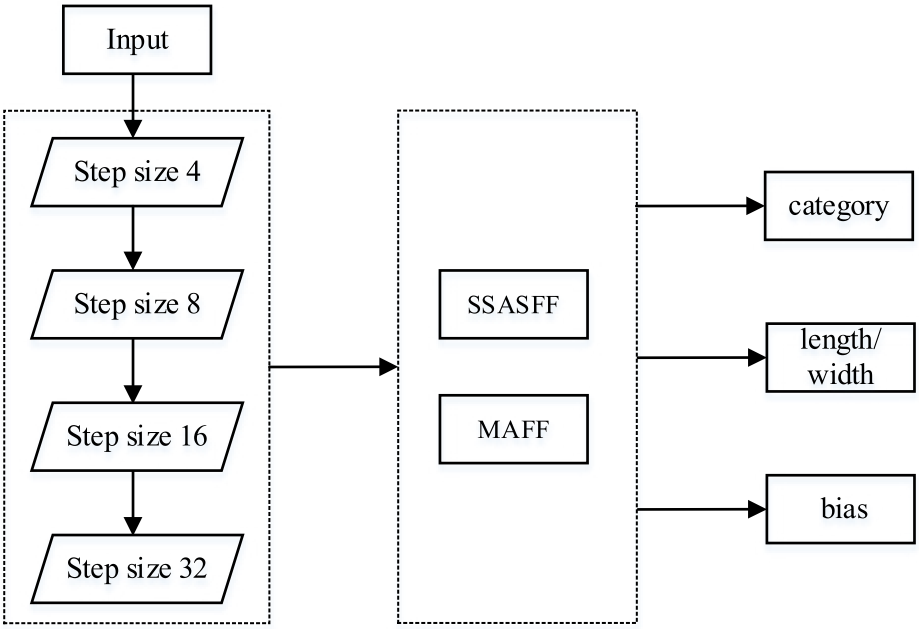


Fig. 1. Structure of MASFF

A. SSASFF

ASFF uses backbone models with low sampling rates, such as DLA and Hourglass23. When using mainstream network structures (e.g. ResNet, Inception, MobileNet), this model only uses 1/4 size of the original image sampled on multiple deformable convolution and deconvolution for prediction. It does not realize the repeated utilization of multilevel features. Therefore, the single-scale adaptive spatial feature fusion module and the multi-level adaptive feature fusion module are proposed to improve the recognition effect.

Adaptive spatial feature fusion (ASFF) was proposed by Li et al. [25] on the basis of YOLOv3 to fuse features of different levels. As shown in Fig. 2, the ASFF module first resamples the features of different scales to the object size (interpolation or deconvolution is adopted for the up-sampling, while pooling is adopted for the down-sampling), and then fuses the output feature maps of three different levels at three different object scales to obtain three feature maps for prediction.

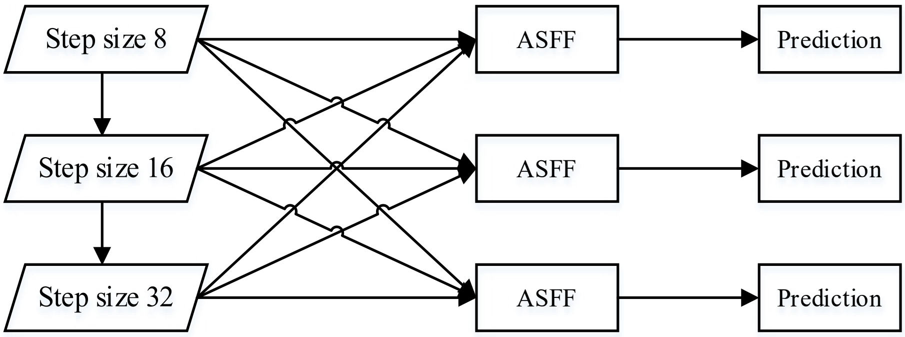


Fig. 2. Structure of ASFF

It can be seen from Fig. 1 and Fig. 2 that the MASFF module and ASFF do not match to each other. ASFF is oriented to multi-scale prediction, while MASFF network can only be single-scale in order to retain their ability to avoid the use of non-maximum suppression (NMS) [26]. Therefore, the single-scale adaptive spatial feature fusion (SASFF) module is proposed in this paper, which is used to fuse multi-level features of the network, as shown in Fig. 3.

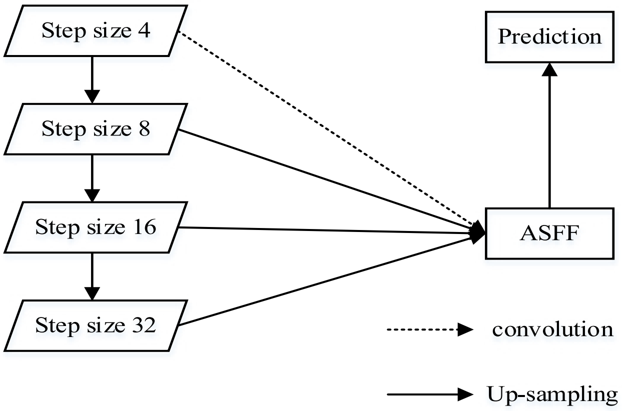


Fig. 3. Structure of SASFF

The proposed single-scale adaptive spatial feature fusion in this paper uses adaptive feature fusion (AFF) module to fuse multi-level features of the network. AFF module firstly adjusts features at different levels into the same size through convolution and up-sampling, and then fuses them with different weights according to their spatial positions. Its expression is:

$$y_{ij} = \sum_{k=1}^{K} a_{ijk} x_{ijk} \quad . \tag{2}$$

Where (i,j) is the spatial coordinate of the feature. K is the number of fusion features. x and y are input and output features respectively. a is the scale factor by adaptive learning, which is used to adjust the feature gravity of different spatial positions. To satisfy $\sum_{k=1}^{K} a_{ij}^{k} = 1$ and $a_{ij}^{k} \in (0,1)$, the scaling factor is defined as:

$$a_{ij}^{k} = \frac{e^{\lambda_{ij}^{k}}}{\sum_{k=1}^{K} e^{\lambda_{ij}^{k}}} \; . \tag{3}$$

λ_{ij}^{k} is given by x^{k} convolved with 1×1.

The SASFF module proposed in this paper not only fuses the multilevel features of the network, but also makes the feature weights in different levels adaptive learning, which effectively improves the recognition effect.

B. MAFF

In order to further improve the real-time performance of the model, a multiscale adaptive feature fusion (MAFF) module is proposed. As shown in Fig. 4, SASFF fuses all input features at one time, but MAFF fuses features at different scales layer by layer through AFF module. The advantages of this module are as follows: feature fusion layer by layer can allow feature transition from deep abstract feature to shallow specific feature. The layer by layer feature fusion avoids the problem of SASFF computing speed decreasing caused by repeated up and down sampling at different scales.

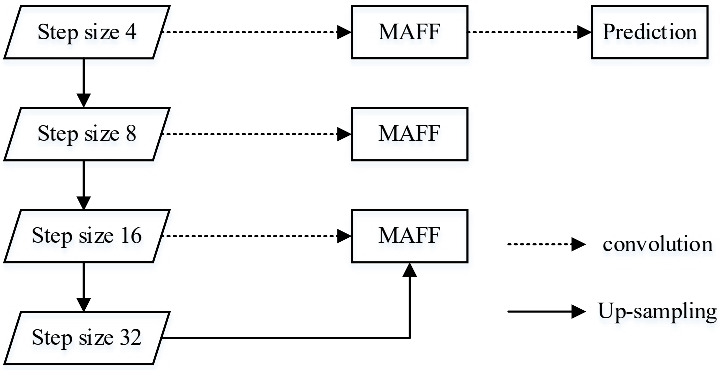


Fig. 4. Structure of MASFF

2.3 Loss Function

The loss based on key point object recognition can be divided into thermal map loss and frame regression loss. Thermal map loss is used for intensive prediction of the object category of the feature map, and the location of the center point of the corresponding category is obtained according to the peak value of the thermal map. The frame regression loss predicts the frame of the object from the center position, so the prediction of the frame depends on the accuracy of the center position to a great extent.

In the frame regression loss, the original ASFF network adopts the error loss function for the width, height and offset of the prediction frame respectively. The original ASFF network assumes that the width, height and offset are independent optimization problems, but in fact the optimization of border is related to each other. In addition, the size of object frame also affects the loss, that is, in the case of the IoU ratio between the same prediction frame and the real frame, the large object tends to achieve greater loss, so that the optimization tends to the large object.

Considering the above problems, the border regression loss in this paper adopts distance IoU (DIoU) loss, and its expression is:

$$L_{DIoU} = 1 - IoU + \frac{d^2}{c^2} \quad . \tag{4}$$

Where d is the distance between the center of the prediction frame and the real frame. c is the diagonal length of the rectangular closure between the prediction box and the real box. The expression for IoU is:

$$IoU = \frac{|B \cap B_{gt}|}{|B \cup B_{gt}|} \quad . \tag{5}$$

B and B_{gt} are prediction boxes and real boxes respectively. It can be seen from equation (4) and Fig. 5 that IoU can constrain the regression of position and length/width of prediction box at the same time, and it is insensitive to the size of object frame. When the prediction box does not overlap with the real box (i.e., IoU=0), d^2/c^2 provides the optimization direction for the loss function.

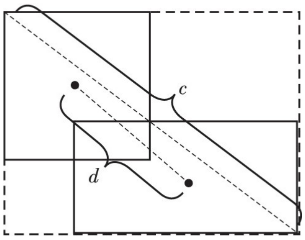


Fig. 5. DIoU loss of predict box and ground truth box

Considering the importance of center position, the paper adopts the variant formula of center loss as the thermal diagram loss, which is expressed as：

$$L_k = -\frac{1}{N}\begin{cases}(1-\hat{Y}_{xyc})^{\alpha}\ln \hat{Y}_{xyc}, & Y_{xyc}=1\\(1-\hat{Y}_{xyc})^{\beta}\hat{Y}_{xyc}^{\alpha}\ln(1-\hat{Y}_{xyc}), & others\end{cases} . \quad \textbf{(6)}$$

Where α =2 and β =4 are hyperparameters. $\hat{Y}_{xyc}$ and Y_{xyc} are predictive confidence and smoothing label, respectively. Y_{xyc} can be expressed as follows:

$$Y_{xyc} = \exp(-\frac{(x-\widetilde{p}_x)^2+(y-\widetilde{p}_y)^2}{2\sigma_p^2}) . \quad \textbf{(7)}$$

Where $\widetilde{p}_x$ and $\widetilde{p}_y$ are the x and y coordinates of the object center point respectively. σ_p is the adaptive standard deviation of the current object.

In this loss function, only the center point of the object is taken as the positive class, while the other positions are set as the negative classes, focusing on the optimization of the center point position. According to formula (6), most of the losses are negative resulting in extremely unbalanced sampling of positive and negative categories. Therefore, points within the radius of the center point are only used as training samples in the training process, and different weights are given to the negative samples. The model avoids using an NMS in its reasoning process. The special setting of the heat map loss results in low confidence at points outside the object center, and the model only predicts at one scale. Therefore, the maximum pooling layer with a size of 3×3 is used to screen out the peak value of the thermal map, and the peak value is taken as the center point of the object recognition.

Therefore, the proposed MASFF in this paper combines thermal diagram loss and frame regression loss. The final loss function is:

$$L = L_k + L_{DIoU} . \quad \textbf{(8)}$$

3 Experiments and Analysis

3.1 Data Set

The data sets used in the experiment are Vehicle model recognition data set, autonomous driving BDD100K data set and object detection Pascal VOC [27] data set. The data set of Vehicle model recognition comes from the actual data collected by cameras in roads, crossings and other places, and part of the data is shown in Fig. 6.

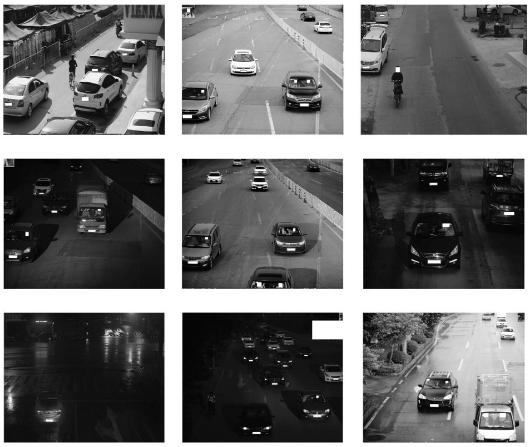


Fig. 6. Some samples in Vehicle data set

There are 10 types of detection objects in Vehicle data set, including forward bus, forward car, forward pickup truck, forward van, forward truck, backward bus, backward car, backward pickup truck, backward van and backward truck. Their categories are represented by 1 to 10, as shown in Fig. 7. In Vehicle data set, training set data contains 26000 images, and test set data contains 6500 images. BDD100K data set is a large-scale and diversified driving video data set, including 10 detection categories, namely, human, bus, cyclist, traffic sign, traffic light, truck, motorcycle, train, bicycle and car. The training set data contains 70000 images, and the verification set contains 10000 images. The Pascal VOC dataset is divided into 20 categories, including planes, bicycles, birds, boats, drinking glasses, buses, cars, cats, chairs, cows, dining tables, dogs, horses, motorcycles, people, potted plants, sheep, sofas, trains, and televisions. There are 16551 images in the training set and 6452 images in the test set. The evaluation indexes adopted in the Vehicle data set are average accuracy (mAP), average accuracy when the IoU is 0.50 (AP50), and average accuracy when the IoU is 0.75 (AP75). AP50 is used on BDD100K and Pascal VOC datasets. The same network model is used to train and test different datasets.

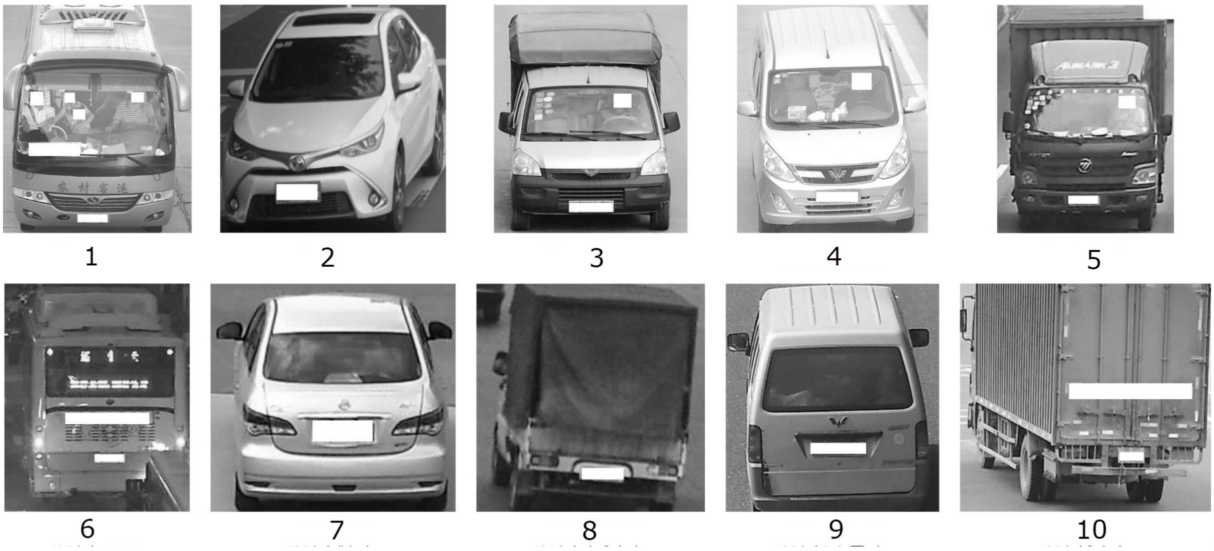


Fig. 7. Images of all classes in Vehicle data set

3.2 Network Parameters Setting

The model training and testing in this paper are conducted in Ubuntu16.04 environment with NVIDIA GTX 1060T graphics card. Python3.7 and Pytorchl.2.0 frameworks are used for training. Network training parameters are as follows: The image input size is 512×512 pixels. The initial learning rate is 5×10^{-4}, and the total number of training is 140. At 90 and 120 training iteration, the learning rate is decreased to 5×10^{-5} and 5×10^{-6}, respectively. The batch size at training time is 64. Adam algorithm is used for optimization, β_1 is 0.9, β_2 is 0.999. During the testing, the batch size is set to 1. Image flipping and multi-scale transformation are not used.

3.3 Experimental Results and Analysis

The Vehicle test results of different models on the Vehicle dataset are shown in Table 1. The size of input images in the training stage is 512×512 pixels. It can be seen from the table that when ResNet18-ASFF model is adopted, the mAP of Vehicle recognition on Vehicle data set is 0.696. When SASFF is adopted, its mAP is 0.705 better than the original ResNet18-ASFF model. SASFF module integrates feature maps of different scales of the network, so the mAP is slightly improved. When the loss function of the original ASFF model is replaced by DIoU loss, the mAP of ResNet18-ASFF-DIoU network is 0.705, which is 0.9 higher than that of ResNet18-ASFF model. It is proved that DIoU loss can effectively improve the effect of vehicle identification. When DIoU loss and feature fusion methods are combined at the same time, the mAP values of network using SASFF and MAFF feature fusion modules are 0.712 and 0.715 respectively, increasing by 1.6 and 1.9% respectively. For YOLOv4 and YOLOv5 models based on candidate boxes, their mAP values are 0.717 and 0.722, respectively, slightly higher than the accuracy of the proposed model, indicating that the effect of the proposed model in vehicle detection and recognition is close to that of YOLOv4 and YOLOv5 models. The size of the proposed model is lower than that of the two models. Compared with EfficientDet-D0 [28] and YOLOv4-tiny [29], the proposed model can extract more features and has obvious advantages in accuracy.

Table 1. Detection results with different methods on vehicle data set

Model	mAP	AP50	AP75	Model size/MB	Video memory/ MB	Time/ms
EfficientDet-D0	0.433	0.554	0.512	15.84	----	12.21
YOLOv4	0.717	0.860	0.827	244.45	1267	18.32
YOLOv4-tiny	0.658	0.852	0.793	22.62	609	2.27
YOLOv5	0.722	0.873	0.828	91.04	1205	13.30
ResNet18-ASFF	0.696	0.854	0.802	55.23	562	5.40
ResNet18-ASFF-DIoU	0.700	0.854	0.806	55.23	562	5.40
SASFF-ASFF	0.705	0.854	0.812	73.83	634	9.07
SASFF-ASFF-DIoU	0.712	0.857	0.816	73.83	634	9.07
MAFF-ASFF-DIoU	0.715	0.861	0.823	55.61	564	6.58

TensorRT6 FP32 model is adopted in this paper. For video memory, the original ResNet18-ASFF model and ResNet18-ASFF-DIoU model have 562MB video memory, indicating that the video memory can be preserved using DIoU loss. When using the feature fusion method, SASFF-ASFF-DIoU and MAFF-ASFF-DIoU models have 634 and 564MB video memory, respectively, indicating that the computing load of SASFF-ASFF-DIoU is significantly increased, while the memory occupied by MAFF-ASFF-DIoU is increased by 2MB, reducing the consumption of computing resources. The video memory of this new model is lower than that of YOLOv4 and YOLOv5 models.

It can also be seen from table 1 that the detection time of the original ResNet18-ASFF model and ResNet18-ASFF-DIoU model is 5.40ms, which indicates that the using of DIoU loss can improve the recognition accuracy while maintaining the video memory and speed. When the feature fusion method is used, the model recognition accuracy is improved obviously, but the speed is decreased. The detection time of SASFF-ASFF-DIoU and MAFF-ASFF-DIoU are 9.07 ms and 6.58ms respectively. The detection time of the proposed model is lower than that of YOLOv4 and YOLOv5 models. EfficientDet-D0 detection time is 12.21ms, the real-time vehicle detection is slower. The detection speed of YOLOv4-tiny is the fastest, but the recognition accuracy is not good enough. The proposed model not only has high accuracy in vehicle detection and recognition, but also has fast speed and good real-time vehicle recognition performance.

The comparison of training loss of each model on the Vehicle dataset is shown in Fig. 8. It can be seen that the training loss of ResNet18-ASFF and ResNet18-ASFF-DIoU model decreases with the increase of training time. The training loss of ResNet18-ASFF-DIoU model decreases faster than that of ResNet18-ASFF model. When the model is stable, the training loss of ResNet18-ASFF-DIoU model is lower than that of ResNet18-ASFF model. When the model is stationary, the model using feature fusion SASFF and MAFF modules has the smallest training loss and high mAP.

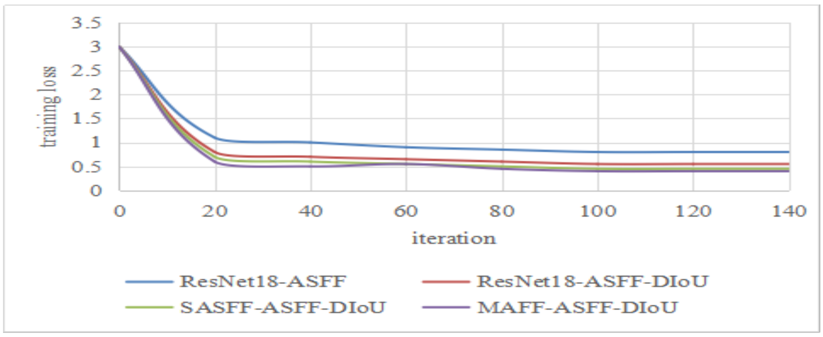
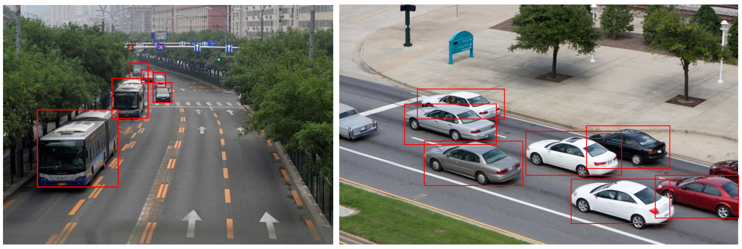


Fig. 8. Training loss with four models

Fig. 9 shows the vehicle detection effect of the SASFF-ASFF-DIoU and MAFF-ASFF-DIoU model. It can be seen that the measured vehicle images can achieve the expected detection effect.



(a) SASFF-ASFF-DIoU (b) MAFF-ASFF-DIoU

Fig. 9. Detection effect of two models

The proposed method in this paper is used to conduct experiments on BDD100K autonomous driving dataset and VOC dataset. The results are shown in Table 2. In the BDD100K dataset, the AP50 of the original ResNet 18-ASFF model is 0.395, which is improved by 0.1% after adding the DIoU loss. After using the feature fusion modules SASFF and MAFF, the AP50 of the model is 0.447, which improves by 5.2% than the original ResNet 18-ASFF model. With the addition of DIoU loss and MAFF modules, the AP50 of MAFF-ASFF-DIoU model is 0.753, which improves by 2.5% than that of ResNet 18-ASFF model on Pascal VOC dataset. It shows that the feature fusion method and the proposed DIoU loss method in this paper can effectively improve the effect of object detection. For BDD100K and VOC datasets, the detection accuracy of YOLOv4 and YOLOv5 is higher than that of the proposed model. The reason is that the new model in this paper mainly solves the problem of vehicle recognition in scenes such as roads and forks, and most of these vehicles are large objects. However, there are many small objects in BDD100K and VOC datasets, so the model in this paper cannot detect and recognize small objects well. The model in this paper adopts ASFF as the backbone network. Compared with CSPNet, the network depth is not enough and the detected objects are not sufficient. Due to the shallow depth of YOLOv4-tiny and EfficientDet-D0 models, they cannot extract more features, so their object detection accuracy is relatively low. In terms of detection time, the results of all models on Vehicle, BDD100K and VOC data sets are basically consistent. The speed of the proposed model in this paper is relatively fast while maintaining high recognition accuracy.

Table 2. Detection results of different models on different data sets

Datasets	Model	AP50	Time/ms
BDD100K	EfficientDet-D0	0.139	12.41
	YOLOv4	0.507	18.54
	YOLOv4-tiny	0.363	2.44
	YOLOv5	0.516	13.34
	ResNet18-ASFF	0.395	5.65
	ResNet18-ASFF-DIoU	0.405	9.29
	SASFF-ASFF	0.403	5.65
	SASFF-ASFF-DIoU	0.447	9.29
	MAFF-ASFF-DIoU	0.447	6.80
VOC	EfficientDet-D0	0.598	12.52
	YOLOv4	0.830	18.59
	YOLOv4-tiny	0.634	2.48
	YOLOv5	0.837	13.37
	ResNet18-ASFF	0.728	5.70
	ResNet18-ASFF-DIoU	0.744	9.38
	SASFF-ASFF	0.743	5.70
	SASFF-ASFF-DIoU	0.751	9.38
	MAFF-ASFF-DIoU	0.753	6.87

4 Conclusion

In order to improve the accuracy of vehicle recognition, a vehicle recognition method based on improved ASFF is proposed in this paper. Firstly, ASFF only uses the last layer of the network features and does not reuse the multilevel features of the network output. In this paper, a single-scale adaptive spatial feature fusion method is proposed to improve ASFF to realize the multilevel features fusion of the network and effectively improve the accuracy of vehicle recognition. Because the single-scale adaptive spatial feature fusion method increases the model size and reduces the reasoning speed, a multiscale adaptive feature fusion method is proposed in this paper, which can effectively improve the model size and reasoning time, and improve the accuracy of vehicle detection and recognition. In terms of the loss function, the length/width loss and offset loss of CenterNet prediction frame are discarded. DIoU loss and thermal diagram loss are used as the loss function of the model, which can improve the convergence speed and recognition accuracy of the model. The detection results on BDD100K and Pascal VOC datasets show that the proposed improved ASFF can significantly improve the vehicle recognition accuracy.

5 Acknowledgement

This paper is supported by Science and Technology Development Plan of Henan Province in 2019. Yuke 2019 Project No. 16. Fund No. 192102210262.

References

[1] J. Mabrouki, M. Azrour, G. Fattah, D. Dhiba, S.-E. Hajjaji, Intelligent Monitoring System for Biogas Detection Based on the Internet of Things: Mohammedia, Morocco City Landfill Case, Big Data Mining and Analytics 4(1)(2021) 10-17.

[2] L. Fang, L. Peng, Design and research on wireless intelligent monitoring system for sewage pipeline leakage of textile mill, Microprocessors and Microsystems 81(4)(2021) 103734.

[3] S.-L. Yin, J. Liu, L. Teng, A new krill herd algorithm based on SVM method for road feature extraction, Journal of Information Hiding and Multimedia Signal Processing 9(4)(2018) 997-1005.

[4] S.-L.Yin, J. Liu, H. Li, A Self-Supervised Learning Method for Shadow Detection in Remote Sensing Imagery, 3D Research 9(4)(2018).

[5] L.-F. Zhang, L.-P. Zhang, D.-C. Tao, X. Huang, A modified stochastic neighbor embedding for multi-feature dimension reduction of remote sensing images, Isprs Journal of Photogrammetry & Remote Sensing 83(2013) 30-39.

[6] A. Ambardekar, M. Nicolescu, G. Bebis, M. Nicolescu, Vehicle classification framework: a comparative study, Eurasip Journal on Image & Video Processing 2014 (1)(2014) 1-13.

[7] L. Briz-Redón, F. Martínez-Ruiz, F. Montes, Identification of differential risk hotspots for collision and vehicle type in a directed linear network, Accident Analysis & Prevention 132 (2019) 105278.

[8] J. Arróspide, M. Camplani, L. Salgado, Image-based on-road vehicle detection using cost-effective Histograms of Oriented Gradients, Journal of Visual Communication and Image Representation 24(7)(2013) 1182-1190.

[9] Y. Yao, B. Tian, F.-Y. Wang, Coupled Multivehicle Detection and Classification With Prior Objectness Measure, IEEE Transactions on Vehicular Technology 66(3)(2017) 1975-1984.

[10]Z. Huo, Y. Xia, B. Zhang, Vehicle type classification and attribute prediction using multi-task RCNN, in: Proc. 2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 2016.

[11]K. Shahid, Y. Zhang, S. Yin, M.-R. Asif, An Efficient Region Proposal Method for Optical Remote Sensing Imagery, in: Proc. IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium, 2018.

[12]S. Yin, Y. Zhang, K. Shahid, Region search based on hybrid convolutional neural network in optical remote sensing images, International Journal of Distributed Sensor Networks 15(5) 2019.

[13]T.-Y. Tang, S.-L. Zhou, Z.-P. Deng, H.-X. Zou, L. Lei, Vehicle Detection in Aerial Images Based on Region Convolutional Neural Networks and Hard Negative Example Mining, Sensors (Basel, Switzerland), 17(2)(2017).

[14]S. Yin, H. Li, L. Teng, Airport Detection Based on Improved Faster RCNN in Large Scale Remote Sensing Images, Sensing and Imaging 21(2020).

[15]X. Wang, A. Shrivastava, A. Gupta, A-Fast-RCNN: Hard Positive Generation via Adversary for Object Detection, in: Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017.

[16]X.-W. Hu, X.-M. Xu, Y.-J. Xiao, H. Chen, S.-F. He, J. Qin, P.-A. Heng, SINet: A Scale-Insensitive Convolutional Neural Network for Fast Vehicle Detection, IEEE Transactions on Intelligent Transportation Systems 20(3)(2019) 1010-1019.

[17]A.-T, Tajar, A. Ramazani, M. Mansoorizadeh, A lightweight Tiny-YOLOv3 vehicle detection approach, Journal of Real-Time Image Processing (5)(2021).

[18]J.-S. Sri, R.-P. Esther, LittleYOLO-SPP: A Delicate Real-Time Vehicle Detection Algorithm, Optik 225 (2020).

X. Wang, S. Wang, J. Cao, Y. Wang, Data-Driven Based Tiny-YOLOv3 Method for Front Vehicle Detection Inducing SPP-Net, IEEE Access 8 (2020) 110227-110236.

[19]G.-M. Cao, X.-M. Xie, W.-Z. Yang, Q. Liao, G.-M. Shi, J.-J. Wu, Feature-Fused SSD: Fast Detection for Small Objects, (2017). arXiv:1709.05054

[20]Y. Chang, P. Chung, H. Lin, Deep learning for object identification in ROS-based mobile robots, in: Proc. 2018 IEEE International Conference on Applied System Invention (ICASI), 2018.

[21]K. Sun, B. Xiao, D. Liu, J. Wang, Deep High-Resolution Representation Learning for Human Pose Estimation, in: Proc.

2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[22]B. Yu, J. Shin, G. Kim, S. Roh, K. Sohn, Non-anchor-based vehicle detection for traffic surveillance using bounding ellipses, (2020). arXiv:2010.02059

[23]N. Chen, H.-D. Xiao, Perceptual audio hashing algorithm based on Zernike moment and maximum-likelihood watermark detection, Digital Signal Processing 23(4)(2013) 1216-1227.

[24]X. Li, W. Li, D. Ren, H. Zhang, M. Wang, W. Zuo, Enhanced Blind Face Restoration With Multi-Exemplar Images and Adaptive Spatial Feature Fusion, in: Proc. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[25]D. Zheng, H. Li and S. Yin, Action Recognition Based on the Modified Two-stream CNN, International Journal of Mathematical Sciences and Computing (IJMSC) 6(6)(2020) 15-23.

[26]M. Everingham, S. Eslami, L.-V. Gool, C.-K.-I. Williams, J. Winn, A. Zisserman, The Pascal Visual Object Classes Challenge: A Retrospective, International Journal of Computer Vision 111(1)(2015) 98-136.

[27]M. Tan, R. Pang, Q.-V. Le, EfficientDet: Scalable and Efficient Object Detection, in: Proc. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[28]A. Bochkovskiy, C.-Y. Wang, H. Liao, YOLOv4: Optimal Speed and Accuracy of Object Detection, 2020. arXiv:2004.10934.