International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 11 Issue 3 ǁ March 2023 ǁ PP. 101-107

A Deep learning-based object detection Method for UAVs

JIA Liyng[1*]
(1. Shandong University of Technology, Zibo, 255000)

Abstract: *To address the problems of existing deep learning-based objection detection methods, which are computationally intensive, take up a lot of memory to run and cannot be deployed on UAV devices, an AI-based intelligent detection and identification method for UAV airborne is proposed. Based on the YOLOv3 object detection model and improved, the model size and number of parameters are reduced without significantly sacrificing detection accuracy. Firstly, a PDW module is proposed based on a depthwise separable convolution and residual network. An object detection model for UAV airborne equipment with shallower network layers and fewer channels is designed. Proposed method finally evaluates on MS-COCO and VisDrone dataset. The experimental results show that the method has a smaller model size, fewer parameters and faster running times than the YOLOv3 network, while maintaining good accuracy, which is significantly better than YOLOv3-tiny, and is also more real-time and accurate than current target recognition detection methods.*

Key words: *Depthwise separable convolution, Object recognition detection, UAVs, Artificial intelligence (AI)*

Date of Submission: 01-03-2023 Date of acceptance: 11-03-2023

I. INTRODUCTION

Since the 21st century, intelligent devices with operational computer vision have a wide range of application space and development prospects. For example, vehicle-mounted intelligent traffic recorder, intelligent extraction robot, intelligent UAV, etc. Among them, as a representative of advanced intelligent equipment, intelligent UAV has been expanding its application market. It is the on-board camera and embedded software system on the UAV that are endowed with computer vision capabilities. The main functions are to identify the types of objects in the scene, locate the positions of these objects, determine the exact boundaries of each object, etc. These scene parsing functions correspond to three basic research tasks in the field of computer vision, namely image classification, detection and recognition, and semantic segmentation. Among them, detection and recognition are the most basic functional module, so object detection has always been a research hotspot in the field of vision [1,2].

Due to the diversity of the open deployment environment, it is difficult to analyze the scene on the front-end equipment of the UAV, which brings many difficulties to the airborne target detection and recognition algorithm. The main difficulties are as follows:(1) How to deal with various changes of visual appearance in aerial detection, such as illumination, Angle, and deformation;(2) How to successfully deploy the target detection and recognition system on the UAV device with limited computing power and memory.(3) How to balance the detection accuracy and real-time requirements.

Object detection and recognition methods based on machine learning and hand-designed features are prone to fail in the face of these changes. At present, the object detection and recognition method based on deep learning seems to be a good way to solve this difficulty.Driven by the advances in computing power of graphics cards and the widespread use of large-scale labeled data sets (e.g., ImageNet and MS-COCO), object detection and recognition methods based on deep learning have become the research focus of more and more researchers due to their good advantages of scalability and end-to-end. At present, there are two kinds of frameworks for object detection algorithms based on deep learning: two-stage detection and recognition frameworks [3-9] and end-to-end one-stage frameworks like YOLO series [10-15]. However, these target detection and recognition methods have large computational overhead and high-power consumption, which are not suitable for intelligent detection and recognition of UAV airborne front-end. Therefore, how to reduce floating-point operations (FLOPs), trainable parameters, and maintain the detection accuracy of object detection and recognition methods without significantly sacrificing detection accuracy is an urgent problem to be solved. At present, there are many ways to "Slim" the object detection model, such as YOLO-tiny, Slim-YOLO, mobile-SSD, etc. [16]. Therefore, this paper mainly develops an algorithm suitable for intelligent detection and recognition of UAV airborne front-end according to the current lightweight target detection and recognition framework. Its main contributions are as follows:(1) A lightweight network unit: PDW architecture is proposed;(2) An intelligent target detection network suitable for UAV airborne front-end is designed.(3) A model framework of intelligent target detection for unmanned aerial vehicle (UAV) is proposed

The remaining chapters of this paper are arranged as follows: Section II introduces related work, Section III details the algorithm suitable for intelligent detection and identification of UAV airborne front-end, Section IV is the comparative experiment part, and Section V concludes.

II. Related Research

2.1 twostage

The two-stage object detection method is also known as the object detection method based on region search, which matches the attention mechanism of the human brain, and first scans the whole image as a whole. Then we find the region of interest (ROI) in the image. In this method, CNN is inserted into the sliding window method, and the topmost feature map is directly predicted according to the position to obtain the results. The RCNN series should be representative of current two-stage object detection methods. Girshick et al. [3] proposed R-CNN neural network and achieved 53.3% Average Precision (mAP) on the PASCAL VOC 2012[4] open dataset. He et al. [5] proposed a new CNN architecture named SPP-net from spatial Pyramid Matching (SPM) in theory. He et al. proposed Fast R-CNN[6] based on SPP-net, which mainly improved the network pooling layer. On the basis of Fast R-CNN, He et al. proposed successively Faster R-CNN[7] and Mask R-CNN[8] algorithms. Among them, Faster R-CNN first proposed the use of Region proposal network for extraction. Compared with R-CNN, the SS(Selective Search) method [9] is used to generate detection boxes, which improves the speed of candidate box generation. Although the two-stage method has high accuracy, it also requires high computing power of the deployed equipment, which is difficult to be applied to the intelligent detection and recognition of the airborne front-end of UAV.

2.2 onestage

The YOLO series algorithm [10-13] is mainly a general object detection model proposed by Joseph et al., which treats the detection problem as a regression problem, improves the detection speed, and accepts input images of different sizes. YOLO, on the other hand, struggles with small objects. To solve this problem, Liu et al. [14] proposed an object detection method named SSD. It is inspired by the Anchor mechanism adopted in MultiBox. (Also refer to the Anchor used in the candidate region recommendation network). Given a specific feature map, there is no fixed grid like the one adopted in YOLO. SSD uses a set of default anchors [15] and uses different aspect ratios and scales to determine the region. In order to deal with objects of different sizes, the SSD network combines the output of multiple feature maps with different resolution sizes. YOLO and SSD series are classic one-stage object detection and recognition frameworks [16]. Although it has a small model size, it cannot be directly deployed on the UAV, so we need to use the model lightweight method to optimize the one-stage object detection framework.

2.3 Model lightweight

Xu et al. [17] further improved based on Inception V3 in 2017, and proposed the depth separable convolution. In the same year, the concept of depth separable convolution was also introduced in Mobilenet, which is a method to miniaturize the network model. The essence is to decompose the standard convolution into two steps. The depthwise separable convolution process is as follows. We use the following formula to calculate and compare the floating-point computation of standard convolution and depthwise separable convolution.
The standard convolution is computed as follows:

$$K_h \times K_w \times C_{in} \times C_{out} \times W \times H \tag{1}$$

The depthwise separable convolution is computed as follows

$$K_h \times K_w \times C_{in} \times W \times H + C_{in} \times C_{out} \times W \times H \tag{2}$$

The ratio of floating-point operations of depthwise separable convolution to standard convolution is:

$$\frac{K_h \times K_w \times C_{in} \times W \times H + C_{in} \times C_{out} \times W \times H}{K_h \times K_w \times C_{in} \times C_{out} \times W \times H} = \frac{1}{C_{out}} + \frac{1}{K_{h,w}^2} \tag{3}$$

According to Equation (3), when the size of the convolution kernel is 3×3, the FLOPs that can be depthwise separated are about 1/9 of the standard convolution. In depthwise separable convolution, it is mainly divided into two steps. First, different convolution kernels are used to convolute different depth channels, and then the standard convolution is decomposed into individual channels. These two processes combined are like the convolutions of standard convolutions, but the model has significantly less computation and model parameters. Therefore, this paper uses the improved depthwise separable convolution when constructing the lightweight and efficient backbone network.

III. Object detection method of UAVs

At present, due to the development of high-performance computing graphics cards and large data set annotation, more and more high-precision object detection models have emerged, but the calculation amount and memory occupancy of the model have also become larger, and they cannot be applied to the airborne front-end of UAV. To this end, this paper designs a new artificial intelligence based neural network model based on YOLO v3-tiny, which has lower computation and higher performance. In this paper, increasing the depth of the network layer enables YOLOv3-tiny to extract richer convolutional features. However, considering that increasing the depth of the network will cause large memory occupation of the network and longer detection time, it is not suitable for the real-time performance required by the airborne front-end of the UAV. Therefore, in this paper, we use the improved depthwise separable convolution and residual module to form the backbone network. FIG. 1 shows the network architecture diagram of the intelligent detection method for the UAV airborne front-end.

As shown in Figure 1, the intelligent detection network framework of UAV in the front-end includes two parts: backbone network and detection network. In line with the goal of network lightweight and enhancing feature extraction ability, we redesigned the basic unit of the backbone network, which is called PDW structure. The backbone network consists of one standard convolution and nine PDW structures. Through this structure, multi-layer feature reuse and fusion can be realized, and the amount of calculation introduced by the new structure can be reduced. In order to detect objects at different distances, the prediction network consists of two branches, corresponding to outputs at two scales. The following article will introduce the difference between PDW structure, standard convolution and depthwise separable convolution.

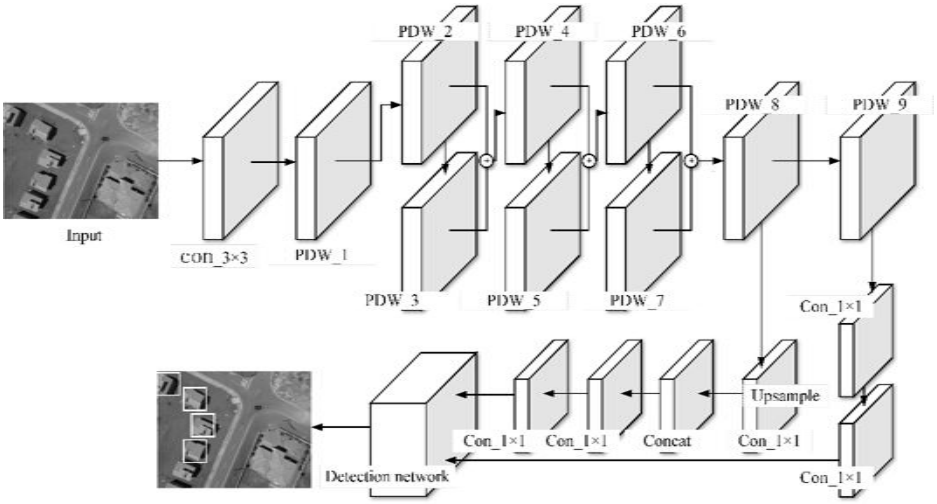

Fig.1 Proposed method model architecture

3.1 PDW structure

The standard convolution is shown in FIG. 2 (a). The addition of BN layer and ReLU in the standard convolution is mainly to speed up the convergence of the model. The BN (Batch normalization) layer is usually placed between the standard convolutional component and the ReLU. The depthwise separable convolution is shown in Figure 2 (b). The improvement over Figure 2 (a) mainly lies in the use of DW convolution and PW convolution. The effect of this structure is like the standard convolution but with much lower computation.

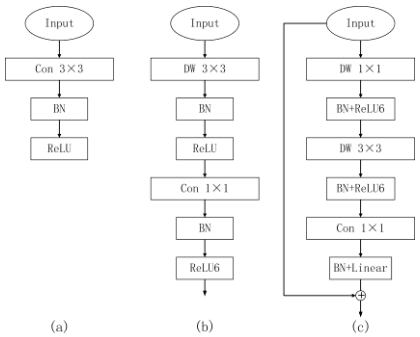

Fig.2 (a) Ordinary convolution; (b) Deep separable convolution (c) PDW structures

FIG. 2 (c) shows the PDW structure proposed in this paper, which mainly adds depthwise separable convolution

to the residual structure. Enhance the propagation of gradients. The improved structure can extract high dimensional feature space. At the same time, this paper adds a BN layer after each convolution operation to speed up the convergence of the model. The output of the first and second steps uses the nonlinear activation function ReLU6 to make the model more robust in low-precision calculations. In the third step, the output of the convolution operation does not use an activation function, but a direct linear output to reduce the loss of information. We use the PDW structure instead of the standard convolution of the original network, which greatly reduces the amount of computation. By increasing the number of channels and network layers, we can improve the accuracy of the model and make it easy to migrate to embedded devices, mobile devices, or UAVs.

3.2 Backbone network

In Figure 1, n in convolutional PDW_n represents the number of times the PDW structure is currently used. The backbone network first performs standard convolution on the input image using a 3×3 convolution kernel to obtain convolution 1, extract features, and upgrade the dimension. In order to enrich the network feature information, we use the idea of residual, add convolutional PDW_2 and convolution_PDW_3 (where convolutional feature maps are added, but the number of channels is unchanged), and then continue to use the PDW structure to generate convolution_PDW_4. Again, Conv_PDw_6 and conv_pdw_8 are obtained. Thus, the backbone network has used the PDW structure nine times. Next, we replace convolution and Max pooling with convolution with a step size of 22. This allows us to keep the number of parameters unchanged and omit the computation involved in Max pooling, which is 1/4 of the original network.

3.3detection Network

The prediction network consists of two branches corresponding to two scales of output. The first branch outputs a 13 × 13 × 18 convolutional feature map. When the convolutional feature map of the backbone reaches Conv_PDW_8, the second branch is generated in parallel. The second branch is fused with the 26 × 26 × 64 output of the first branch, resulting in a final output of 26 × 26 × 18. We used the k-means clustering method to calculate six anchors for the current dataset: (50×66), (74×99), (91×125), (113×154), (140×190), and (220×284). K-means uses the Euclidean distance. In the backbone network output, the first three anchors are used for 13×13 convolutional feature maps, and the last three anchors are used for 26×26 convolutional feature maps. For 13×13 and 26×26 outputs, each parameter includes x, y, w, h, confidence, and probability.

3.4Loss functions

The loss function of the proposed algorithm consists of four parts. The first part involves the prediction of the center coordinates, as shown in Equation (4). The second part involves the regression prediction of the bounding box, as shown in Equation (5); The third part involves the prediction of the object category, as shown in Equation (6). The fourth part concerns the prediction of object confidence, as shown in Equation (7).

$$Loss_1 = \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{obj} \left[\left(x_i - \hat{x}_i \right)^2 + \left(y_i - \hat{y}_i \right)^2 \right]$$

where, $x_i$and y_i are the coordinates of the predicted object, x_i*and* $y_i$are the coordinates of the true value.

$$Loss_2 = \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \delta_{ij}^{object} \left[\left(\sqrt{w_i} - \sqrt{\hat{w}_i} \right)^2 + \left(\sqrt{h_i} - \sqrt{\hat{h}_i} \right)^2 \right]$$

where, w_i and h_i are the length and width of the predicted object, w_i and $h_i$are the length and width of the true value.

$$Loss_3 = \sum_{i=0}^{S^2} \delta_{ij}^{object} \sum_{j=0}^{B} \left[p_i(C) - \hat{p}_i(C)^2 \right]$$

so, the total loss is:

$$Loss_{total} = \sum_{i=0}^{S^2} Loss_i$$

3.5 Algorithm flow

The main process of artificial intelligence-based UAV detection algorithm includes two parts: training and test. As shown in Figure 6, model training requires input of the data set and labels to the constitutional neural network; The training is then done alliteratively and a trained neural network is built. In the test, the information of the image obtained by the UAV's on-board camera is input into the training model, and the position of the target in the image is obtained for recognition and detection.

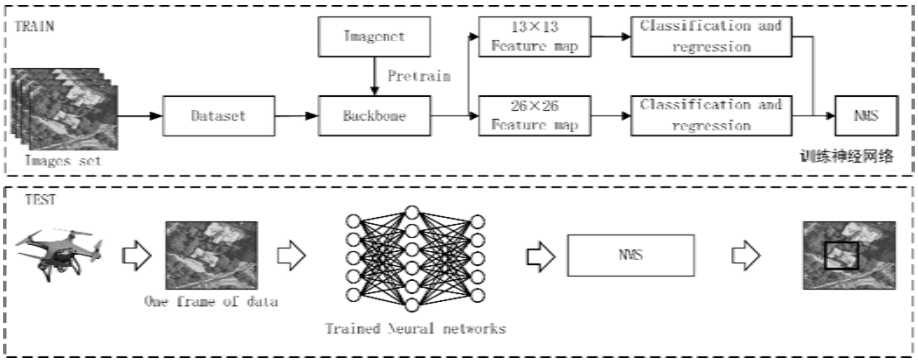

Fig.3 Algorithmic flow

IV. Experiment and analysis

4.1 Experimental setting

We used many environments and tools for image recognition and detection in the experiment, as shown in Table 1:

Tab 1 Tools used in vehicle detection algorithms

Projects	Details
Operating system	Linux Ubuntu 18.04
Computing architecture	CORE i7-9750h + NVIDIA GTX 1080Ti
Framework	Pytorch
Image processing framework	Python 3.7、OpenCV 3.0

4.2 Datasets

Microsoft Common Objects in Context (MS-COCO) dataset:MS-COCO dataset is an image classification, image detection, and image segmentation dataset built by Microsoft team. The dataset has the following characteristics: (1) Object independence (2) Context recognition (3) multiple objects per image (4) Over 300,000 images (5) over 2 million instances (6) 80 object categories.

VisDrone dataset: The VisDrone dataset was collected by the AISKYEYE team at the Machine Learning and Data Mining Laboratory of Tianjin University. The benchmark dataset consists of 400 video clips consisting of 265,228 frames and 10,209 still images.

4.3 Experimental results of MS-COCO

In Table 2, the proposed method is compared with the more advanced object recognition and detection methods that currently exist. All methods are trained with MS-COCO dataset. The comparison information in the table includes the backbone network, accuracy, and run time (ms), which are calculated on the CPU. The FPS and time of the two-stage method are not discussed since they are much lower than those of the one-stage method.

Tab 2 detection results on the MS-COCO data set

	Methods	Backbone	Avg. Precision, IOU			Time
			0.5:0.95	0.5	0.75	
Two-stage	Fast R-CNN	VGG-16	20.5	39.9	19.4	2000
	Faster R-CNN	VGG-16	21.9	42.7	-	142ms
	Faster R-CNN	ResNet-101	27.2	48.4	-	200ms
	R-FCN	ResNet-101	29.9	51.9	-	111ms
	Faster R-CNN w FPN	Res101-FPN	36.2	59.1	39.0	167ms
	Mask R-CNN	ResNeXt-101	39.8	62.3	43.4	303ms
	Fitness-NMS	ResNet-101	41.8	60.9	44.9	200ms
	Cascade R-CNN	Res101-FPN	42.8	62.1	46.3	143ms
One-stage	SSD300	VGG-16	25.1	43.1	25.8	23ms
	SSD512	VGG-16	28.8	48.5	30.3	45ms
	DSSD321	ResNet-101	28.0	46.1	29.2	105ms
	RetinaNet400	ResNet-101	31.9	49.5	34.1	81ms

RefineDet320	VGG-16	29.4	49.2	31.3	25ms
Corner Net	Hourglass	40.5	57.8	45.3	227ms
YOLOv2	DarkNet-19	21.6	44.0	19.2	15ms
YOLOv3 608	DarkNet-53	33.0	57.9	34.4	50ms
YOLOv3 416	DarkNet-53	31.0	55.3	31.7	29ms
Proposed method	Proposednet	28.7	53.4	29.9	13ms

Table 2 shows that the speed of the one-stage target recognition detection is significantly faster than that of the two-stage target recognition detection method. The proposed method is more than two times faster than YOLOv3 416, but the accuracy is slightly reduced. Therefore, the use of PDW structure can significantly reduce the amount of calculation and speed up the operation time of the model, and at the same time can ensure the stability of the accuracy. It can also be seen in the table that the speed of SSD300, RefineDet320, and YOLOv2 can be comparable to that of the proposed method, but their accuracy is lower. Although the accuracy of Mask R-CNN, Cascade R-CNN, Corner Net and other methods is higher than that of the proposed method, the speed of the proposed method is several times faster than that of them

4.3Experimental results of UAVS

To ensure the practicability of our algorithm, YOLOv3-tiny is selected as the comparison algorithm in this paper on the UAVs data set, and YOLOv3-tiny object detection algorithm is applied to UAV front-end detection and recognition to compare with the algorithm proposed in this paper. In this experiment, we set that only when the IOU between the detected target and the real labeled target is greater than 70 percent, the detection and recognition task is completed, otherwise it is still treated as failure, and then the recall rate and accuracy rate are calculated to obtain the accurate F-measure. The experimental results are shown in Figure 10:

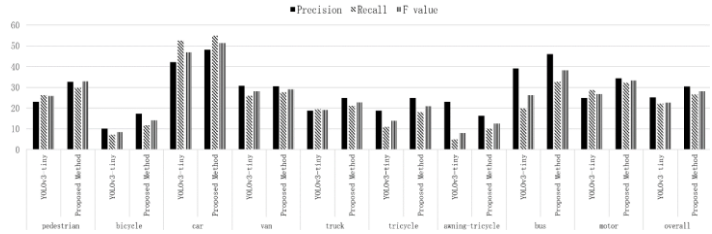

Fig.10 Performance of the YOLOv3-tiny algorithm and proposed method detection result

compared with the YOLOv3-tiny method, the proposed method has improved the precision recall rate and F value under various categories. From the comprehensive results, the F value of the proposed method is 5.6 percentage points higher than that of the YOLOv3-tiny algorithm. In general, the proposed method has higher accuracy than the original YOLOv3-tiny method, and has higher accuracy in recognition and detection of this data set.

V. Conclusion

Due to the large amount of calculation, model parameters and memory consumption of the current target detection method based on deep learning, it cannot be successfully deployed to the UAV for target detection tasks. To solve this problem, this paper takes YOLOv3 as the base algorithm and improves it, and proposes a UAV detection method based on artificial intelligence. In this paper, the main contributions are as follows:1) A PDW structure is proposed to reduce the problem of large amount of data and high calculation of traditional convolution structure;2) An improved network structure based on YOLOv3 target recognition and detection network is given.3) The proposed method is verified on different data sets and compared with the existing methods. It can be obtained that the proposed method has higher real-time performance and accuracy than the current target recognition and detection methods.

References

[1]. Pato, L.V., Negrinho, R.M., & Aguiar, P.M. (2019). Seeing without Looking: Contextual Rescoring of Object Detections for AP Maximization. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 14598-14606.

[2]. Ancha, S., Nan, J., & Held, D. (2019). Combining Deep Learning and Verification for Precise Object Instance Detection. ArXiv, abs/1912.12270.

[3]. GirshickR , Donahue J , Darrelland T , et al. Rich feature hierarchies for object detection and semantic segmentation[C]// 2014 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2014.

[4]. Everingham M ,Gool L V , Williams C K I , et al. The Pascal Visual Object Classes (VOC) Challenge[J]. International Journal of Computer Vision, 2010, 88(2):303-338.

[5]. PurkaitP , Zhao C , Zach C . SPP-Net: Deep Absolute Pose Regression with Synthetic Views[C]// British Machine Vision

Conference(BMVC 2018). 2017.
[6]. GirshickR . Fast R-CNN[J]. Computer Science, 2015.
[7]. Ren S , He K , Girshick R , et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 39(6):1137-1149.
[8]. KaimingH , Georgia G , Piotr D , et al. Mask R-CNN[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, PP:1-1.Uijlings J R R , K. E. A. van de Sande…. Selective Search for Object Recognition[J]. International Journal of Computer Vision, 2013, 104(2):154-171.
[9]. Redmon J ,Divvala S , Girshick R , et al. You Only Look Once: Unified, Real-Time Object Detection[C]// Computer Vision & Pattern Recognition. IEEE, 2016.
[10]. Redmon J , Farhadi A . YOLO9000: Better, Faster, Stronger[C]// IEEE Conference on Computer Vision & Pattern Recognition. IEEE, 2017:6517-6525.
[11]. TianchiJ ,Qiang L I , Maosong L , et al. Target detection method combining inverted residual block and YOLOv3[J]. Transducer and Microsystem Technologies, 2019.
[12]. BochkovskiyA , Wang C Y , Liao H Y M . YOLOv4: Optimal Speed and Accuracy of Object Detection[J]. 2020.
[13]. OlteanG ,Florea C , Orghidan R , et al. Towards Real Time Vehicle Counting using YOLO-Tiny and Fast Motion Estimation[C]// 2019 IEEE 25th International Symposium for Design and Technology in Electronic Packaging (SIITME). IEEE, 2019.
[14]. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C., & Berg, A. SSD: Single Shot MultiBox Detector. ECCV, 2016.
[15]. Zhang P , Zhong Y , Li X . SlimYOLOv3: Narrower, Faster and Better for Real-Time UAV Applications[C]// 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). IEEE, 2020.
[16]. Sandler M , Howard A , Zhu M , et al. MobileNetV2: Inverted Residuals and Linear Bottlenecks[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2018.
[17]. CholletF .Xception: Deep Learning with Depthwise Separable Convolutions[J]. arXiv e-prints, 2016.