Journal of Highway and Transportation Research and Development
2024, 18(3): 48−60

Research Article
ISSN 2095−6215
DOI: 10.26599/HTRD.2024.9480022

BaAM-YOLO: a balanced feature fusion and attention mechanism based vehicle detection network in aerial images

Xunxun Zhang[1*], Xu Zhu[2]

1. *Xi'an University of Architecture & Technology, Xi'an Shaanxi 710055, China*;
2. *Chang'an University, Xi'an Shaanxi 710064, China*

Abstract: Vehicle detection in aerial imagery is of paramount importance for various intelligent transportation applications, including vehicle tracking, traffic control, and traffic behavior analysis. However, this task presents significant challenges due to factors such as scale variance, the presence of small vehicles, and complex scenes conditions. The demand for high-quality and precise vehicle detection exacerbates these challenges, particularly in complex environments characterized by low light and occlusion. To address these issues, we propose a balanced feature fusion and attention mechanism-based vehicle detection network, termed BaAM-YOLO, specifically designed for aerial images. Firstly, to improve the vehicle detection accuracy, we didn't directly fuse deep and shallow features. Instead, we focused on achieving a balanced contribution from different feature layers. Secondly, to enhance the feature extraction capabilities of our model, we incorporated a simple parameter-free attention mechanism (SimAM), which aims to improve detection accuracy. Additionally, to effectively model the target confidence loss and classification loss, we emplogyed focal loss to derive a loss function that facilitates dynamic adjustment for multi-class classification. The proposed vehicle detection method was evaluated using the VisDrone2019 dataset. Comparative analyses with state-of-the-art research indicate that the proposed method can obtain superior and competitive results in vehicle detection.

Keywords: traffic engineering; vehicle detection; balanced feature fusion; attention mechanism; focal loss

1 Introduction

Unmanned aerial vehicle (UAV) is playing an increasingly important role in the intelligent transport system (ITS), including vehicle tracking[1], event detection[2], and behavior analysis[3]. Compared with fixed cameras, UAV has a greater advantage in traffic information gathering and urban road network planning, which mainly profits from the broader perspective and the low cost. Besides, it can take off and land in any empty area and is less affected by environmental factors. Currently, vehicle detection in aerial images has drawn great attentions[4-5], but it is still challenging due to the variable sizes of the vehicles, the complex background, and with occlusion. Specially, for vehicle detection at night or with occlusion and high density, it becomes confronted with more and more challenges. Hence, there is an urgent need for an adaptive vehicle detection method in aerial images in the complex scenes.

The vehicle detection in aerial images becomes more challenging than in fixed images, and gets more attention from domestic and foreign scholars. Due to the particularity of aerial imaging view and long imaging distance, the aerial images contain more and larger complex scenes than the ground images. Therefore, it leads to plenty of tiny vehicles and serious scale variance. Nowadays, the frequently-used vehicle detection methods are based on deep learning such as CenterNet[6], Faster Region-Convolutional Neural Networks (R-CNN)[7], and Cascade R-CNN[8].

Deep learning-based vehicle detection methods can be divided into two categories: two-stage detection and one-stage detection. Two-stage vehicle detection needs to generate a large number of candidate boxes on the feature map in advance, then perform high-quality box screening, and finally classify and position regression,

Received: 8 October 2023; **Revised:** 25 May 2024; **Accepted:** 6 June 2024

* *E-mail address*: zxx@xauat.edu.cn

while single-stage target detection completes classification and regression in one step. The two-stage object detection algorithm is represented by the R-CNN series[9-10], while the single-stage classic algorithm is represented by Single Shot MultiBox Detector (SSD)[11], You Only Look Once (YOLO) [12], and RetinaNet[13].

Compared with the two-stage vehicle detection algorithm, the single-stage algorithm presents a relatively ideal effect. Therein, the YOLO series algorithms receive the most attention due to their high speed and excellent detection effect. Representatively, Redmon et al. proposed the YOLO network based on the regression of target position[12]. Compared to the current two-stage methods, the YOLO network is with less accuracy and a lower recall rate. To overcome these problems, Redmon and Farhadi constructed the YOLO9000 network[14], but it is not satisfactory for small targets. To improve the detection accuracy for small targets, Redmon et al. put forward the YOLOv3 network via incremental improvement[15]. To fully utilize the huge number of features and improve Convolutional Neural Network (CNN) accuracy, Bochkovskiy et al. use new features including Weighted-Residual-Connections (WRC), Cross-Stage-Partial-connections (CSP), Cross mini-Batch Normalization (CmBN), Self-adversarial-training (SAT), and Mish-activation to achieve state-of-the-art results[16]. Furthermore, Glenn Jocher et al. tinker with the YOLOv4 network and construct the YOLOv5 network, which adds some new improvement ideas (including Mosaic data enhancement at the input end, adaptive anchor box calculation, and adaptive image scaling operation) to YOLOv4 and greatly improves the detection speed and accuracy[17].

Considering the diverse requirements for speed and accuracy in the real environment, Li et al. integrate the up-to-date object detection advancements either from industry or academia and build a suite of deployment-ready networks at various scales to accommodate diversified use cases[18]. To meet the requirements of real-time vehicle detection, Wang et al. propose a trainable bag-of-freebies oriented solution to construct the YOLOv7 network[19]. It combines the flexible and efficient training tools with the proposed architecture and the compound scaling method. Due to its high precision and real-time performance, the YOLOv7 network is a perfect candidate for vehicle detection in aerial images.

However, in most practical scenarios, complex scenes such as foggy weather, rainy days, weak light, and occlusion inevitably cause varying degrees of false detection rate and missing detection rate. To fully extract global and local feature information, Lu et al. propose a high-quality object detection method, which constructs a hybrid backbone based on the combination of convolutional neural network (CNN) and vision transformer (ViT)[20]. Hu et al. propose a high-precision detection method based on an improved faster region-based convolutional neural network (Faster-RCNN), which incorporates a feature pyramid structure to the backbone network to effectively extract the features[21]. These methods improve the detection accuracy based on deep feature extraction. However, it may reduce the speed of the training and make the network to be more complex. Besides, the previous studies focus on enhancing the connection of the different feature layers, but little consideration on the balance of the portion of the feature layers.

Besides, for the loss function of the YOLOv7 network, it is applied to update the loss of the gradient, which is composed of the coordinate loss, the target confidence loss, and the classification loss[19]. Actually, there is a significant difference in the quantity of the samples in the practical scenarios. To conquer the imbalance problem of samples, we utilize the binary to the make the dynamic adjustment of the focal loss for the weight. For the vehicles in the aerial images, we divide the samples into different types. Therefore, we utilize the focal loss to deduce the loss function of the multi-classification dynamic adjustment.

Based on the analysis above, we propose a balanced feature fusion and attention mechanism based vehicle detection network in aerial images. To improve the accuracy of vehicle detection, we fuse the deep and shallow features. The deep feature is utilized to provide rich semantic information, while the shallow feature focuses on specific and spatial information. Firstly, to improve the vehicle detection accuracy, we don't directly fuse the deep and shallow features. Alternatively, we give more attention on the balance of the portion for the different feature layers. Secondly, to enhance the capability of the proposed feature extraction, we introduced the simple parameter-free attention mechanism (SimAM) to improve the vehicle detection accuracy. Finally, to charac-

terize the target confidence loss and the classification, we utilize the focal loss to deduce the loss function of the multi-classification dynamic adjustment.

In a word, the main contributions of our work are two-fold. Firstly, to improve the precision of the vehicle detection, we not only fuse the deep and shallow feature layers, but also consider the balance of the portion of the different layers. It is also a main measure to improve the vehicle detection ability in the complex scene. Secondly, to characterize the target confidence loss and the classification, we utilize the focal loss to deduce the loss function of the multi-classification dynamic adjustment, which can not only improve the accuracy of the detection results but also obtain a more accurate fused bounding box of the target vehicles.

The thesis can be divided into four parts. The introduction is given in Section 1. Section 2 presents the methodology of the balanced feature fusion and attention mechanism based vehicle detection network in aerial images. Experiments and analysis are provided in Section 3. Then, Section 4 summarizes the conclusion.

2 Methodology

In this part, we first construct a balanced feature fusion and attention mechanism based vehicle detection network in aerial images. Then, the balanced linear transform feature fusion is elaborated, which not only fuses the deep and shallow layers, but also considers the portion of the feature layers. In the following, the loss function for multi-class classification tasks is detailed.

2.1 Vehicle detection network in aerial images

In this work, we select the YOLOv7 network as our basic network due to its high accuracy and computing speed. It is a one-stage object detection network, which is widely applied in real-time detection due to its high computing speed. Especially, the accuracy and speed are superior to the other series of YOLO networks.

For the YOLOv7 network, there are three layers: input layer, backbone layer, and head layer. The input layer is applied to pre-process the input image and then input it to the backbone layer. The backbone layer mainly focuses on extracting features with three different sizes, which consists of 51 convolutional modules. Besides, the backbone layer adopts the efficient E-ELAN network architecture, leading to high detection efficiency. For the head layer, it aims to confuse the features from the backbone layer and generalize the bounding boxes to predict the label, which outputs the predicted results of three different sizes.

As illustrated in Figure 1, we construct a balanced feature fusion and attention mechanism based vehicle detection network in aerial images. To improve the accuracy of vehicle detection, we fuse the deep and shallow features in proportion. The deep feature is utilized to provide rich semantic information, while the shallow feature focuses on specific and spatial information. Secondly, to enhance the capability of the proposed feature extraction, we introduced the simple parameter-free attention mechanism (SimAM) to improve the vehicle detection accuracy. Furthermore, to detect the tiny vehicles in aerial images, we construct a multi-scale aerial vehicle detection strategy, which adds a scale of 160× 160 to detect the tiny vehicles. Finally, to characterize the target confidence loss and the classification, we utilize the focal loss to deduce the loss function of the multi-classification dynamic adjustment. Besides, the MP-Conv and SPPCSPC modules are presented in Figure 2.

Based on the analysis above, we propose a balanced feature fusion and attention mechanism based vehicle detection network in aerial images. Compared with the original YOLOv7 network, we have made three improvements. Firstly, to improve the vehicle detection accuracy, we don't directly fuse the deep and shallow features. Alternatively, we give more attention on the balance of the portion for the different feature layers. Secondly, to enhance the capability of the proposed feature extraction, we introduced the SimAM module to improve the vehicle detection accuracy. Finally, to characterize the target confidence loss and the classification, we utilize the focal loss to deduce the loss function of the multi-classification dynamic adjustment. In the following, the balanced linear transform feature fusion and the loss function for multi-classification tasks are presented.

2.2 Balanced linear transform feature fusion

Due to the various and complicated sizes of the vehicles in the aerial images, it is extremely difficult to accurately detect the vehicles. Besides, the distance between the vehicle and the camera could significantly influence the size of the vehicle, leading to worse vehicle detection accuracy. Therefore, it is extremely difficult to

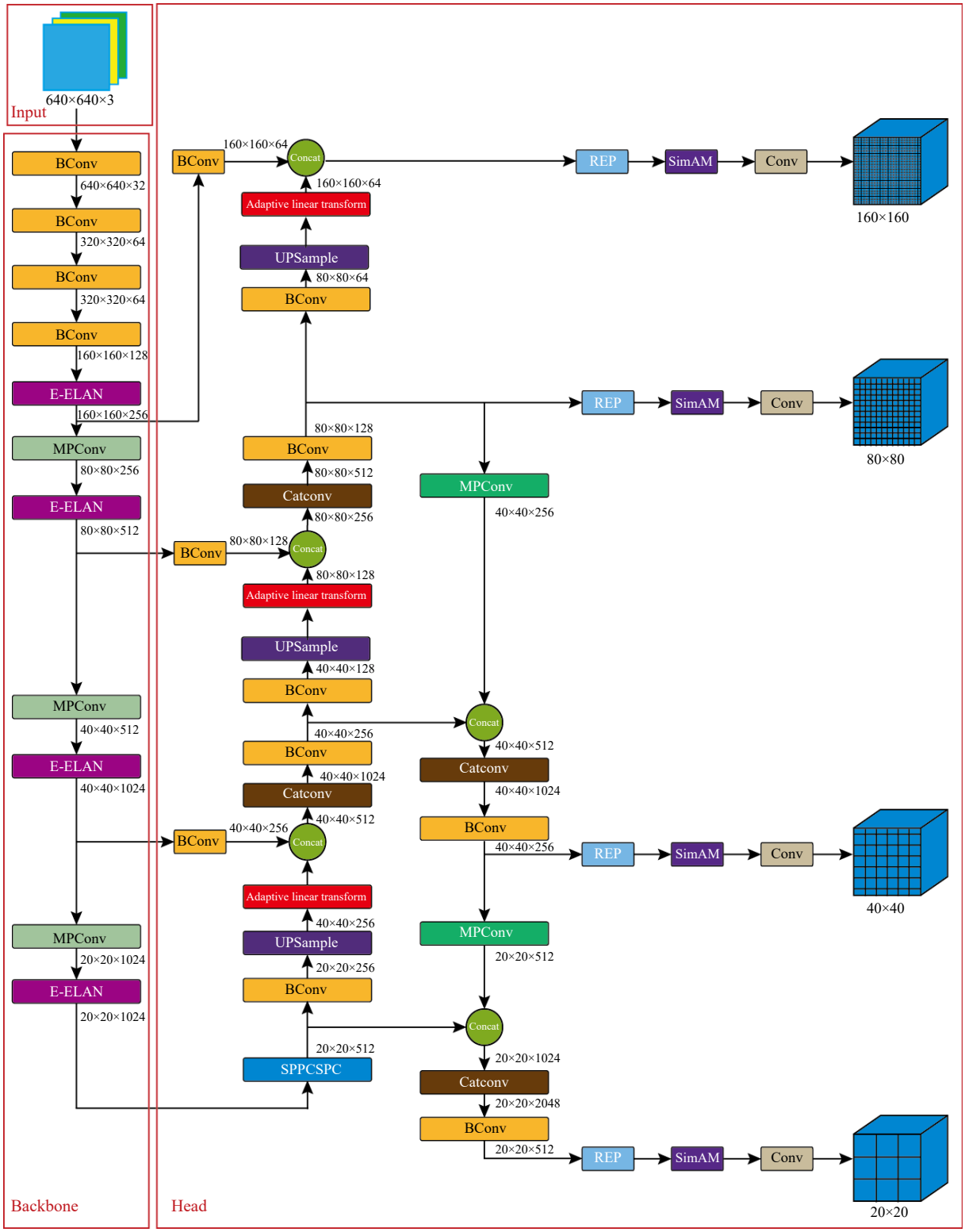


Fig. 1 The structure of the proposed vehicle detection network in aerial images

accurately detect the vehicles using only a single feature layer.

To conquer this problem, we fuse the deep and shallow features to improve vehicle detection accuracy. In the feature map, the deep and shallow features contain different information. For the deep feature, it has a big receptive field and focuses on the global information, which has rich semantic information. For the shallow feature, its receptive field is smaller, and it focuses on the specific information and has rich spatial information.

The original YOLOv7 network adopts the FPN structure, which utilizes different sizes of features to perform

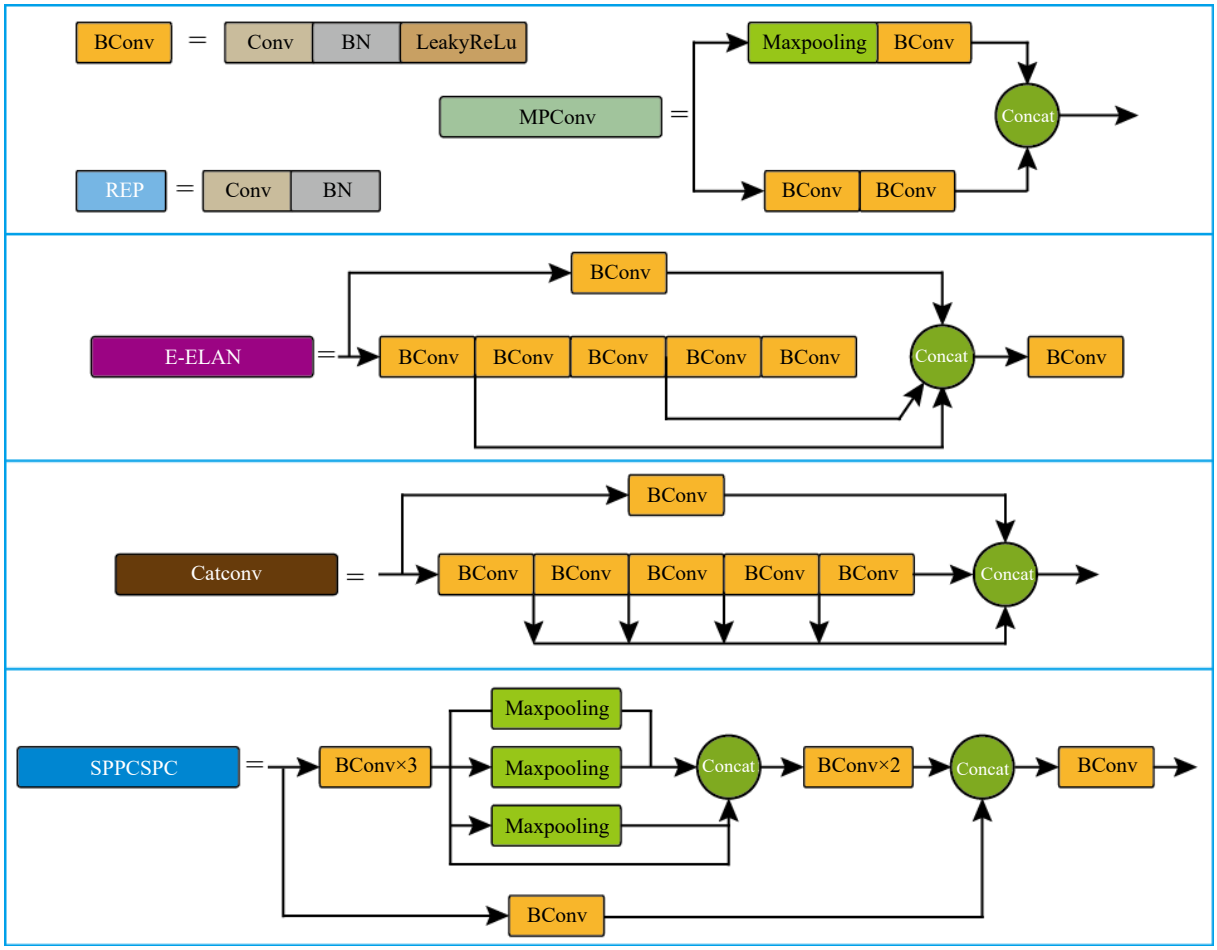


Fig. 2 The MPConv and SPPCSPC modules of the vehicle detection network

the parallel prediction. It fuses the deep and shallow features to detect the vehicle of small size using the high-definition shallow feature, and detect the vehicle of big size using the low-definition shallow feature. However, the proportion of the advanced Semantic information and spatial geography information is not considered in the FPN structure, and the FPN structure only fuses the deep feature after up-sampling, which cannot supply sufficient semantic information for the shallow feature. Therefore, in our work, we add the balanced linear transform module after the up-sampling. The balanced linear transform module first takes a linear transform of the deep feature, and then adjusts the weight of the deep feature to merge with the shallow feature. It can balance the portion of the advanced Semantic information and spatial geography information to improve the detection accuracy of small vehicles.

As illustrated in Figure 1, the process of the balanced linear transform feature fusion is presented. Therein, the convolution of the feature map (the output of the SPPCSPC module) with the size of 20×20×512 is performed to obtain the feature map with the size of 20×20×256. After up-sampling, the feature map with the size of 40×40×256 is sent to the balanced linear transform module. Then, the output of the balanced linear transform module and the original feature map with the size of 40×40×256 are combined.

For the balanced linear transform feature, it can be expressed as:

$$\boldsymbol{D} = [\boldsymbol{M} * \boldsymbol{K} + \boldsymbol{b}, \boldsymbol{O}] \tag{1}$$

where, $\boldsymbol{D}$ represents the fused feature of the deep and shallow features; $\boldsymbol{M}$ and $\boldsymbol{O}$ denote the deep and shallow feature, respectively; $[\cdot, \cdot]$ is the splicing operation; $\boldsymbol{K}$ is the matrix with the same dimension of the matrix $\boldsymbol{D}$, while $\boldsymbol{b}$ is the corresponding bias.

Similarly, the feature map with the size of 80×80×128 is performed. To sum up, by introducing the balanced linear transform feature fusion module, the proportion of the deep and shallow features is considered. Therefore, the balanced linear transform feature fusion module is conducive to the detection of vehicles with various and

complicated sizes in aerial images.

For the baseline YOLOv7 network, there are three scales of detection: 80×80, 40×40, and 20×20, respectively. In the aerial images, the size of the target vehicles is relatively small. In our work, there are target vehicles with a size of less than 8×8. To conquer this problem, we introduce an extra 160×160 prediction layer.

2.3 SimAM module

To enhance the capability of the proposed feature extraction, we introduce the simple parameter-free attention mechanism (SimAM), which differs from the traditional channel or spatial attention modules. The SimAM module provides a more efficient way to accomplish the image processing tasks, which does not require additional parameters to compute the attention weights of feature map. The SimAM module can not only enable better capture of fine-grained details but also improve the vehicle detection precision.

Owing to the SimAM module, the proposed vehicle detection method can focus more on the crucial features that are relevant to vehicles, resulting in the high detection accuracy. This enhancement can enable the vehicle detection method to better handle severe challenges, such as occlusion, variations in viewpoint, and changes in lighting conditions.

The SimAM module extracts fundamental features based on the energy function, and the energy function of each neuron is given as follows:

$$e_t(w_t, b_t, y, x_i) = \frac{1}{M-1}\sum_{i=1}^{M-1} [-1-(w_t x_i + b_t)]^2 + [1-(w_t x_i + b_t)]^2 + \lambda w_t^2 \tag{2}$$

where, t and x_i represent the target neurons and other neurons of input feature $X \in \mathrm{R}^{C\times H\times W}$; i is an index for spatial dimensions; $M = H \times W$ represents the number of neurons in each channel; w_t and b_t are given as follows:

$$\begin{cases} w_t = -\dfrac{2(t-\mu_t)}{(t-\mu_t)^2 + \dfrac{2}{M-1}\displaystyle\sum_{i=1}^{M-1}(x_i-\mu_t)^2 + 2\lambda} \\ b_t = -\dfrac{1}{2}(t+\mu_t)w_t \\ \mu_t = \dfrac{1}{M-1}\displaystyle\sum_{i=1}^{M-1} x_i \end{cases} \tag{3}$$

According to Equations 2 and 3, we can obtain the minimum energy function as follows:

$$e_t^* = \frac{4\left[\dfrac{1}{M}\displaystyle\sum_{i=1}^{M}\left(x_i - \frac{1}{M}\sum_{i=1}^{M}x_i\right)^2 + \lambda\right]}{\left(t - \dfrac{1}{M}\displaystyle\sum_{i=1}^{M}x_i\right)^2 + \dfrac{2}{M}\displaystyle\sum_{i=1}^{M}\left(x_i - \frac{1}{M}\sum_{i=1}^{M}x_i\right)^2 + 2\lambda} \tag{4}$$

Then, the SimAM module can be described as:

$$X = \mathrm{sigmoid}\left(\frac{1}{E}\right)\odot X \tag{5}$$

where E groups $\frac{1}{e_t^*}$ in the channel, and spatial dimensions and sigmoid are used to prevent weights from becoming excessively large.

2.4 Loss Function for Multi classification tasks

The loss function of the YOLOv7 network is applied to update the loss of the gradient, which is composed of the coordinate loss L_{ciou}, the target confidence loss L_{obj} and the classification loss L_{cls}. It can be expressed as:

$$L_{\mathrm{loss}} = L_{\mathrm{ciou}} + L_{\mathrm{obj}} + L_{\mathrm{cls}} \tag{6}$$

Therein, we adopt the logarithmic binary cross entropy loss to characterize the target confidence loss L_{obj} and the classification loss L_{cls}.

To conquer the imbalance problem of samples, we utilize the binary to make the dynamic adjustment of the focal loss for the weight. For the vehicles in the aerial images, we divide the samples into three types: vehicles under good light conditions, vehicles at night with weak light, and vehicles with occlusion. Therefore, we utilize the focal loss to deduce the loss function of the multi-classification dynamic adjustment.

For the multi-classification task, the softmax function is selected as the activation function. Therein, we set the three labels as $l_1(0, 0, 1)$, $l_2(0, 1, 0)$, and $l_3(1, 0, 0)$. Then, the output of the softmax is (p_1, p_2, p_3). p_1, p_2, and p_3 represent the probability of the corresponding category, respectively. Besides, $p_1 + p_2 + p_3 = 1$. The Focal Loss is designed to address the one-stage object detection where there is an extreme imbalance between foreground and background classes during training. We introduce the focal loss starting from the cross entropy loss for binary classification:

$$L_{\mathrm{ce}}(p, y) = L_{\mathrm{ce}}(p_t) = -\lg(p_t) \tag{7}$$

where, $y \in 1, -1$ specifies the ground-truth class, while $y \in [0, 1]$ denotes the probability of the corresponding class with the label as $y = 1$; p_t is an intermediate vari-

able. Besides, $p_t = p$ if $y = 1$, otherwise, $p_t = 1 - p$.

To address the class imbalance, we introduce a weighting factor $\lambda \in [0, 1]$ for $y = 1$ and $1 - \lambda$ for $y = -1$. Similarly, we define η_t as we define p_t. Then, the η-balanced cross entropy loss is:

$$L_{\eta-\mathrm{ce}}(p_t) = -\eta_t \lg(p_t) \tag{8}$$

Although η can balance the importance of positive and negative samples, it does not differentiate between easy and hard samples. Instead, we propose to reshape the loss function and thus focus training on hard negatives. Furthermore, we add a modulating factor $(1 - p_t)^{\gamma}$ to the cross entropy loss. Then, the focal loss is defined as:

$$L_{\mathrm{fl}}(p_t) = -(1 - p_t)^{\gamma} \lg(p_t) \tag{9}$$

Analogously, we utilize the η-balanced variant of the focal loss as the cross entropy loss:

$$L_{\eta-\mathrm{fl}}(p_t) = -\eta_t (1 - p_t)^{\gamma} \lg(p_t) \tag{10}$$

So far, Equation (10) is the multi-class cross entropy. To balance the weight of the positive and negative samples, we add η_t to Equation (9). Furthermore, to reduce the portion of the samples that are easy to classify, we introduce $(1 - p_t)^{\gamma}$ as the attenuation. In this work, the label is an one shot. Therefore, the value of the label is 1 in the corresponding position, otherwise, the value is 0. Finally, we combine the balance factor η_t and the attenuation $(1 - p_t)^{\gamma}$ to obtain the final balanced focal loss as Equation (10). For the attenuation factor γ, we obtain the best value by comparing the results of the simulation. Besides, the attenuation factor γ is interactive with the balance factor η_t. Moreover, the attenuation factor γ has a greater impact than the balance factor η_t.

3 Experiments and analysis

3.1 Training data sets and experimental environment

To verify the effectiveness and efficiency of the proposed vehicle detection approach, we build the network framework illustrated in Figure 1. Therein, we test our proposed vehicle detection in aerial images on the data set VisDrone2019[22]. This data set consists of 288 video clips, i.e. 10209 aerial images and 261908 frames. These images and videos are captured by different unmanned aerial vehicle(UAV) cameras with different heights and angles. Besides, these aerial images include many scenarios: general detection scenario, detection at night, detection with occlusion, and so on. Noticeably, these aerial images are captured in different conditions to meet all the demands of the experimental verification.

To accomplish the experiments and verification, we set the experimental environment and parameters as shown in Table 1.

Table 1 Experimental environment and parameters setting of the proposed framework

Experimental environment		Parameters setting	
System Environment	Ubuntu18.04	Weights	YOLOv7.pt
GPU	RTX 3070Ti	hyp	hyp.scratch.p5.yaml
Development Environment	Pytorch1.10	Epochs	20
Programming Language	Python3.8.13	Batch-size	16
Programming Platform	Anaconda 4.12.0, CUDA11.3	Workers	8

3.2 The effect of Loss Function on the vehicle detection

According to Equation 10, the parameter η has an important influence on the model. In this work, we utilize the *loss.py* file to change the parameter f1-γ of the dataset configuration file. Therein, is a intermediate variable in the dataset configuration file. We set three sample parameter weights as 0.3, 0.3, and 0.4, and select the parameter $\eta \in [0, 1.5]$ to make the comparisons with the parameter $\eta = 0$. As shown in Table 2, for the training loss, the higher the parameter η the lower the training loss. Even worse, $\eta = 0.5$ leads to nonconvergence of the model. For the mean average precision, when $0.5 \leqslant \eta \leqslant 1.5$, mAP has been improved. For the precision, it may reduce when $\eta < 0.5$ or $\eta > 1.5$. For the recall, it increases when $\eta \geqslant 0.5$.

Table 2 Different attenuation factors and the corresponding evaluation results.

Attenuation parameter	Precision	Recall	mAP
0	78.0	83.7	81.5
0.5	74.6	82.9	81.6
1.0	82.2	83.9	85.2
1.5	80.9	80.0	82.6

To sum up, when $0.5 \leqslant \eta \leqslant 1.5$, the improved loss function can bring great improvement. According to Table 2, mAP of the improved loss function is improved

by 3.7% when $\eta = 1.0$. Therefore, we select η as 1.0 in the following work.

3.3 Vehicle Detention Performance of the Proposed Method

To test and verify the vehicle detection effect of the proposed method, we train the proposed network on the VisDrone data set. As shown in Figure 3, the partial vehicle detection results in the aerial images are presented, where the detection results are consistent with the location and classification. In the first row in Figure 3(b), the majority of the vehicles are accurately detected, which is due in large part to the combination of balanced feature fusion and loss function for multi-classification tasks. Besides, the aerial images in the first row are under good light conditions. In the second row, the aerial images are captured at night. The weak light causes an adverse impact on vehicle detection. According to Figure 3(b), there are a small amount of missed and false detections. For the third row in the Figure 3(c), some vehicles are not detected or misdetected. However, for the second and third rows, satisfactory vehicle detection results have still been achieved, despite weak light and occlusion. This is because the weak light or occlusion has a bad effect on the feature extraction.

As shown in Figure 3, we classify and display the vehicle detection results under different circumstances: under good light conditions, at night with weak light, and with occlusion. To further verify the superiority of the proposed method, we compare it with the other YOLO series of methods. As illustrated in Figure 4, we compare the detection results with the other three methods, i.e., YOLOv3, YOLOv5l, and YOLOv7. According to Figure 4, the proposed method presents satisfactory and superior performances. Besides, the confidence in the vehicle detection for the YOLOv7 network and the proposed method is higher than in the YOLOv3 and YOLOv5l networks. For vehicle detection under good light conditions, the four methods all show satisfactory results. However, for vehicle detection at night with weak light, there is false detection in the YOLOv3 network, while there are missed detections in the YOLOv5l and YOLOv7 networks. Similarly, for the vehicle detection with occlusion, there are false and missed detections in YOLOv3, YOLOv5l, and YOLOv7 networks. The proposed method presents satisfactory and superior performances, which mainly gives credit to the fusion of the deep and shallow features in proportion.

Furthermore, to perform a quantitative analysis of the proposed method, we present FPS (Frame Per Second) and mAP (mean Average Precision) to make the comparisons with the other YOLO series of methods. As shown in Table 3, we compare the detection results with the oth-

(a) Under the good light condition

(b) At night with light

(c) With occlusion

Fig. 3 Vehicle detection performance of the proposed method in the different scenes.

(a) YOLOv3

(b) YOLOv5

(c) YOLOv7

(d) The proposed method

Fig. 4 Comparisons of the vehicle detection performance of the different methods.

er five methods, i.e., YOLOv3, YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv7. Therein, sAP is for the average precision of the single category. According to Table 3, the average precision of the proposed network is significantly higher than the other networks, especially for the average precision of the single category. This needs to give the credit to the effort of balanced linear transform feature fusion and loss function for multi-classification tasks. Nevertheless, within the same number of iterations, the FPS of the proposed network is slightly lower than YOLOv5s, YOLOv5m, and YOLOv7 networks. This is because a number of parameters has been added, however, this detection speed can meet the demands of vehicle detection.

Besides, according to Table 3, YOLOv5s shows the best performance on FPS, which is attributed to its light-weight network structure with a small quantity of parameters. While the proposed method presents excellent performance on the mAP, which owes to the feature fu-

Table 3 Performance and comparisons with the other YOLO series of methods.

Model	FPS	mAP	AP of the vehicle detection		
			Under the good light condition	At night	With occlusion
YOLOv3[15]	47.4	78.9	83.1	77.4	76.2
YOLOv5s[17]	94.8	74.1	78.6	72.1	71.6
YOLOv5m[17]	86.6	73.8	75.3	73.8	72.3
YOLOv5l[17]	55.7	76.6	77.8	75.2	76.8
YOLOv7[19]	56.8	81.3	87.3	79.1	77.5
Proposed	55.1	87.5	93.4	85.3	83.8

sion and its reasonable loss function. Compared with YOLOv5s, YOLOv5m, and YOLOv5l, the mAP of the proposed method is increased by 18.1%, 18.6%, and 14.2%.

In a word, we classify and display the vehicle detection results under different circumstances: under good light conditions, at night, and with occlusion. For the special scenes, we also compare the vehicle detection results with the other networks in Figure 4. According to Table 3, AP of the corresponding single category is presented. For the vehicle detection results in the general scene, the AP of the proposed method is increased by 7.6% compared with the YOLOv7 network. However, the proposed method shows superior performance in vehicle detection at night and with occlusion. The corresponding AP is increased by 9.1% and 8.1%. It is demonstrated that the proposed method can improve the precision of vehicle detection, which can effectively conquer the bad effect of weak light and occlusion on the vehicle detection.

3.4 Ablation experiment on the proposed vehicle detection in aerial infra-red images

To verify the effectiveness of the proposed vehicle detection method, we design the ablation experiments on the datasets for validation. In this study, we adjust the structure of the deep neural network to balance the detection accuracy and running time. Therein, we fuse the deep and shallow features, and give more attention to the balance of the portion for the different feature layers. Secondly, we increase the scales from three to four, which is conducive to the detection of the small vehicles in the aerial images. In addition, we add the SimAM module to enhance the network's feature extraction capability. Finally, to characterize the target confidence loss and the classification, we utilize the focal loss to deduce the loss function of the multi-classification dynamic adjustment. Therefore, we divide the ablation experiment into seven groups, and the corresponding experiment configurations are presented in Table 4. The group 7 is the proposed vehicle detection method.

To perform a quantitative analysis of the ten groups, we present the values of the evaluation indicators for different groups in Table 5. Therein, we choose $mAP_{50\%}$, precision, recall, and FPS for comparison. According to Table 5, the group 7 achieves the best effect. By comparing the group 7 with groups 2 and 3, it can be seen that the balanced feature fusion is conducive to the improvement of $mAP_{50\%}$, precision, and recall. For group 7 with group 4, it is indicated that the SimAM module plays a significant role on the improvement of the vehicle detection. For group 7 with group 5, $mAP_{50\%}$, the precision and recall of group 6 are improved because of the four scales. However, FPS is slightly decreased owing to the added scale. For group 7 with groups 6, the multi-classification dynamic adjustment is valuable to improve the detection accuracy.

Table 4 Experiment configurations of the seven groups in the ablation experiment.

Group	Feature fusion	Balanced feature fusion	SimAM	Scale	Loss function
1				3	CIOU
2	√		√	4	Proposed
3			√	4	Proposed
4		√		4	Proposed
5		√	√	3	Proposed
6		√	√	4	CIOU
7		√	√	4	Proposed

Table 5 Evaluation indicators for seven groups.

Group	$mAP_{50\%}$	Precision	Recall	FPS
1	0.412	0.908	0.716	52.3
2	0.446	0.942	0.752	54.3
3	0.424	0.927	0.711	55.6
4	0.431	0.934	0.733	54.6
5	0.437	0.928	0.731	53.9
6	0.461	0.941	0.758	54.4
7	0.467	0.957	0.761	56.3

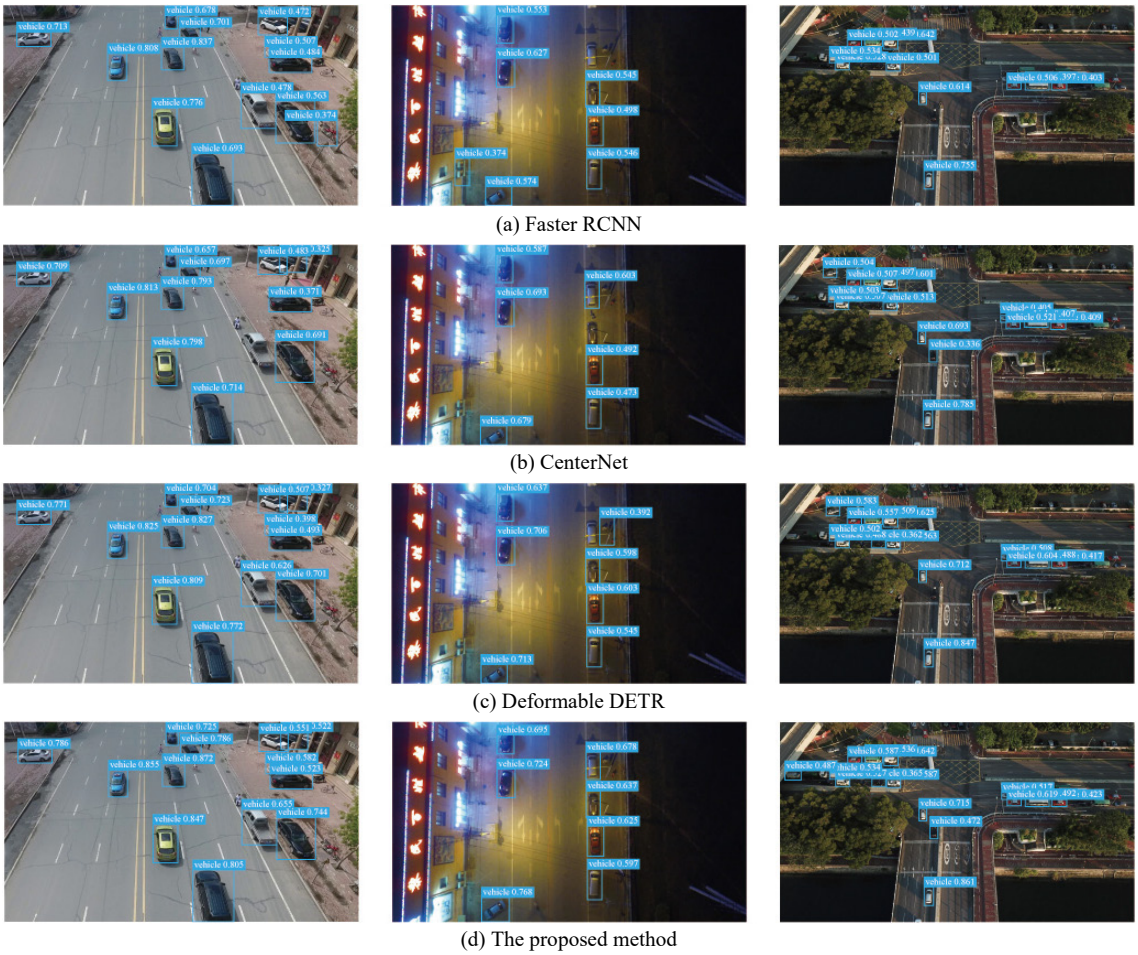
According to Tables 4 and 5, the proposed method presents the satisfactory results. The balanced feature fusion, the SimAM module, four-scale, the multi-classification loss function all contribute to the improvement of the BaAM-YOLO network.

3.5 Comparisons of the vehicle detection performance with the other kinds of methods

In part 3.2, we present not only the vehicle detection performance but also the comparisons with the other five series YOLO methods. In this part, to further verify the superiority of the proposed method, we compare the

vehicle detection results with the other kinds of methods: Faster RCNN, CenterNet, and Deformable DETR[23].

As illustrated in Figure 5, we compare the detection results with the other three methods. According to Figure 5, the proposed method presents satisfactory and superior performances. Besides, the confidence of the vehicle detection for the proposed method is higher than the other three networks. For vehicle detection under good light conditions, the four methods all show satisfactory results. However, for the vehicle detection at night with weak light, there is false detection in the Faster RCNN network, while there are missed detections in CenterNet and Deformable DETR networks. Similiarly, for the vehicle detection with occlusion, there are false and missed detections in the other three networks.



(a) Faster RCNN

(b) CenterNet

(c) Deformable DETR

(d) The proposed method

Fig. 5 Vehicle detection performance of the proposed method in the different scenes.

Furthermore, to perform a quantitative analysis of the proposed method, we also compare FPS and mAP to make the comparisons with the other three methods in Table 6. Therein, the average precision of the proposed network is significantly higher than the other networks, especially for the average precision of the single category. According to Table 6, within the same number of iterations, Deformable DETR shows the best FPS. The proposed method is slightly lower than the CenterNet network. However, the proposed method shows the best performance on mAP and AP of the single vehicle detection. Compared with Faster RCNN, CenterNet, and Deformable DETR, the mAP of the proposed method is increased by 11.3%, 7.6%, and 18.1%. Especially, for the vehicle detection at night with weak light, the mAP of the proposed method is increased by 21.5%, 9.2%, and 20.7%. While for the vehicle detection with occlusion, the mAP is increased by 12.0%, 9.5%, and 14.9%.

According to the comparisons with the other methods in Table 6, the proposed shows excellent and superior performances on vehicle detection. It is demonstrated that the proposed method can improve the precision of

Table 6 Comparision with the other methods.

Model	FPS	mAP	AP of the vehicle detection		
			Under the good light condition	At night	With occlusion
Faster RCNN[7]	47.4	76.9	83.9	70.2	74.8
CenterNet[6]	56.8	81.3	89.3	78.1	76.5
Deformable DETR[23]	84.8	74.1	78.7	70.7	72.9
Proposed	55.1	87.5	93.4	85.3	83.8

vehicle detection, which can effectively conquer the bad effect of weak light and occlusion on vehicle detection.

4 Conclusion

With the goal of overcoming the challenges encountered in the vehicle detection in aerial images, such as the scale variance, small vehicles, and complex scenes, we proposed a balanced feature fusion and attention mechanism based vehicle detection network in aerial images. In this study, we constructed a BaAM-YOLO network to quickly and accurately detect vehicles in aerial images. Rather than directly fusing the deep and shallow features, we gave more attention to the balance of the portion for the different feature layers. In addition, we introduced the SimAM module to improve the vehicle detection accuracy. Finally, experimental results on the VisDrone2019 data set showed that the proposed BaAM-YOLO network can effectively and efficiently detect vehicles with low missed and false detection rates. Compared with the faster R-CNN, CenterNet, Deformable DETR, and a series of YOLO neural networks, the proposed network presented satisfactory and competitive results in terms of the detection accuracy and time required. Furthermore, vehicle detection experiments under different environments were carried out and showed that our method achieved an excellent performance in terms of the detection accuracy and robustness of the vehicle detection.

References

[1] Mao, X. Y.; Yu, B. G.; Shi, Y. J.; et al. Real-Time online trajectory planning and guidance for terminal area energy management of unmanned aerial vehicle[J]. Journal of Industrial & Management Optimization, 2023, 19(3): 1945-1962.

[2] Abdullah, A.; Constantine, E.; Frederic, B. The detection of clandestine graves in an arid environment using thermal imaging deployed from an unmanned aerial vehicle[J]. Journal of Forensic Sciences, 2023, 68(4): 1286-1291.

[3] Zhao, X. Y.; Wang, P.; Zhu, S. Y.; et al. Spatial distribution of traffic conflicts in interchange merging area based on video recognition[J]. Journal of Highway and Transportation, Research and Development, 2021, 38(5): 90-99.

[4] Ma, Y.; Zou, L. P.; Zhang, L. Q. Application technology of uav equipped with optical camera and digital infrared imagery in inspecting quality of pylon and stay cable of sea-crossing bridge[J]. Journal of Highway and Transportation Research and Development, 2018, 35(8): 89-105.

[5] Li, X. W.; Shi, L. X.; Tang, J. Q.; et al. Determinants of passengers' ticketing channel choice in rail transit systems: New evidence of e-payment behaviors from Xi'an, China[J]. Transport Policy, 2023, 140(1): 30-41.

[6] Huang, P. C.; Liu, S.; Zou, Z.; et al. Helmet wearing detection method based on improved centernet with enhanced associations[J]. Computer Engineering and Applications, 2023, 59(17): 250-256.

[7] Niyigena, G.; Lee, S. J.; Kwon, S. K.; et al. Real-time detection and classification of scirtothrips dorsalis on fruit crops with smartphone-based deep learning system: preliminary results[J]. Insects, 2023, 14(6): 523-536.

[8] Wen, X. P.; Lai, H. C.; Gao, G. X.; et al. Video abnormal behaviour detection based on Pseudo-3D encoder and multi-cascade memory mechanism[J]. IET image processing, 2023, 17(3): 709-721.

[9] Sun, P.; Wang, W. Y.; Chai, Y. N.; et al. RSN: Range sparse net for efficient, accurate lidar 3D object detection[J]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021: 5721-5730.

[10] Zhang Y. J.; Yin, Y.; Shao, Z. Y. An enhanced target detection algorithm for maritime search and rescue based on aerial images[J]. Remote Sensing, 2023, 15(19): 4818-4829.

[11] Liu, C. G.; Cheng, H. D.; Zhang, Y. T.; et al. Robust multiple cue Fusion-Based High-Speed and nonrigid object tracking algorithm for short track speed skating[J]. Journal of Electronic Imaging, 2016, 25(1): 013014.

[12] Redmon, J.; Divvala, S.; Girshick, R.; et al. You Only Look

Once: Unified, Real-Time object detection[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas: IEEE Press, 2016: 779-788.

[13] Yue, B. Y.; Chen, L.; Shi, H.; et al. Ship detection in sar images based on improved retinanet[J]. Journal of Signal Processing, 2022, 38(1): 128-136.

[14] Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Hawaii: IEEE Press, 2017: 7263-7271.

[15] Redmon, J.; Farhadi, A. YOLOv3: An incremental improvement[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Salt Lake City: IEEE Press, 2018: 1-7.

[16] Bochkovskiy, A.; Wang, C. Y.; Liao, H. Y. M. YOLOv4: Optimal speed and accuracy of object detection[C]//IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Seattle: IEEE Press, 2020: 1-10.

[17] Jocher, G.; Chaurasia, A.; Stoken, A.; et al. ultralytics/yolov5: v7. 0-yolov5 sota realtime instance segmentation. Zenodo, 2022zndo. . . 7347926J[CP/OL]. DOI:10.5281/zenodo.7347926.2022.

[18] Li, C. Y.; Li, L. L.; Jiang, H. L.; et al. YOLOv6: A Single-Stage object detection framework for industrial applications[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE Press, 2022: 1-17.

[19] Wang, C. Y.; Bochkovskiy, A.; Liao, H. Y. M. YOLOv7: Trainable Bag-Of-Freebies sets new State-Of-The-Art for Real-Time object detectors[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE Press, 2023: 7464-7475.

[20] Lu, W. J.; Niu, C. Y.; Lan, C. Z.; et al. High-Quality object detection method for uav images based on improved dino and masked image modeling[J]. Remote Sensing, 2023, 15(19): 4740-4751.

[21] Hu, Z. C.; Liu, J.; Jiang, C.; et al. A high-precision detection method for coated fuel particles based on improved faster Region-Based convolutional neural network[J]. Computers in Industry, 2022, 143.

[22] Sun, Y. M.; Cao, B.; Zhu, P. F.; et al. Drone-Based RGB-Infrared Cross-Modality vehicle detection via uncertainty-aware learning[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(10): 6700-6713.

[23] Wen, F.; Wang, M.; Hu, X. J. DFAM-DETR: Deformable feature based attention mechanism DETR on slender object detection[J]. IEICE Transactions on Information and Systems, 2023, E106. D(3): 401-409.