Computers, Materials & Continua
DOI: 10.32604/cmc.2022.029239
Article

Tech Science Press

An Algorithm for Target Detection of Engineering Vehicles Based on Improved CenterNet

Pingping Yu[1], Hongda Wang[1], Xiaodong Zhao[1,*] and Guangchen Ruan[2]

[1]School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang, 050000, China
[2]Indiana University at Bloomington, Indiana, 47405, USA
*Corresponding Author: Xiaodong Zhao. Email: zhaoxiaodong@hebust.edu.cn
Received: 28 February 2022; Accepted: 06 May 2022

Abstract: Aiming at the problems of low target image resolution, insufficient target feature extraction, low detection accuracy and poor real time in remote engineering vehicle detection, an improved CenterNet target detection model is proposed in this paper. Firstly, EfficientNet-B0 with Efficient Channel Attention (ECA) module is used as the basic network, which increases the quality and speed of feature extraction and reduces the number of model parameters. Then, the proposed Adaptive Fusion Bidirectional Feature Pyramid Network (AF-BiFPN) module is applied to fuse the features of different feature layers. Furthermore, the feature information of engineering vehicle targets is added by making full use of the high-level semantic and low-level fine-grained feature information of the target, which overcomes the problem that the original CenterNet network did not perform well in small target detection and improve the detection accuracy of the network. Finally, the tag coding strategy and bounding box regression method of CenterNet are optimized by introducing positioning quality loss. The accuracy of target prediction is increased by joint prediction of center position and target size. Experimental results show that the mean Average Precision (mAP) of the improved CenterNet model is 94.74% on the engineering vehicle dataset, and the detection rate is 29 FPS. Compared with the original CenterNet model based on ResNet-18, the detection accuracy of this model is improved by 16.29%, the detection speed is increased by 9 FPS, and the memory usage is reduced by 43 MB. Compared with YOLOv3 and YOLOv4, the mAP of this model is improved by 19.9% and 5.61% respectively. The proposed method can detect engineering vehicles more quickly and accurately in far distance. It has obvious advantages in target detection compared with traditional methods.

Keywords: Engineering vehicle; target detection; CenterNet; feature fusion

1 Introduction

In the power system, the overhead high voltage transmission line is the lifeblood of the whole system, which is related to the safety and benefit of the whole power system, so it plays a crucial role in ensuring the stable operation. At present, the anti-external force of overhead high-voltage transmission line video monitoring is mainly carried out by manual [1]. The monitoring personnel look for possible external intrusion targets by monitoring the static images returned from the front site, such as excavators, cranes, bulldozers and other engineering vehicles. If external intrusion behavior is found, an alarm will be issued and power grid maintenance personnel will be sent to the site to deal with it. As the power grid size growing, limited staffing cannot achieve real-time monitoring of the line environment, therefore, there is an urgent need for an intrusion detection method which can automatically process images and judge the intrusion target to relieve the manual pressure.

At present, for target detection of engineering vehicles, there are mainly traditional detection methods and deep learning based detection methods. Traditional detection methods detect vehicle targets through image edge feature, color feature and Histogram of Oriented Gradient (HOG) [2,3] feature combined with Support Vector Machine (SVM) [4]. However, traditional methods cannot effectively extract target semantic information and have high computational cost, which makes it difficult to adapt to complex scenes around transmission lines. With the success of deep learning in image recognition [5], engineering vehicle detection based on deep learning has gradually become a hot research issue. Li et al. [6] constructed a two-stage vehicle target detection algorithm based on Faster R-CNN [7] by improving RPN network, thus improving the detection accuracy. Zhang et al. [8] used faster-RCNN model to detect the engineering vehicles in the images taken by UAV, realizing the detection of the engineering vehicles in the case of long-distance and small targets, but the model has a large amount of calculation and poor real time. RCNN series models are two-stage target detection models [9], with the development of one-stage detection models, more attention has been paid to the detection of engineering vehicles using them. Yu et al. [10] improved the accuracy of vehicle detection by improving SSD model. Yan et al. [11] improved YOLOv2 target detection algorithm to achieve vehicle target detection. Pu et al. [12] improved the loss function and Non-Maximum Suppression (NMS) part of YOLOv3 algorithm to improve the detection accuracy of engineering vehicles under the condition of loss of certain detection efficiency, but there was a problem of insufficient extraction of semantic features of engineering vehicles. Zhang et al. [13] achieved target detection of excavators in aerial photography environment by using YOLOv4 algorithm. Zhang et al. [14] proposed a target detection method based on YOLOv5 to solve the problem of small and difficult target detection in aerial images, which improved detection accuracy and recall rate. Zhang et al. [15] proposed a vehicle re-identification model based on optimized DenseNet121 with joint loss, which improved the accuracy of vehicle detection.

In conclusion, the existing deep learning-based methods for detecting engineering vehicles mostly adopt the anchor method, but this method is prone to the problem of wrong selection of anchor when detecting targets with different scales. At the same time, the feature extraction of targets cannot make full use of semantic information among feature layers, which will lead to the omission of small targets [16,17]. Therefore, aiming at the problems of insufficient target feature extraction, low detection accuracy and poor real time in long-distance and small-scale engineering vehicle detection, this paper proposes a method based on improved CenterNet for detecting engineering vehicles.

The paper is organized as follows: The second section describes the traditional CenterNet target detection model. The third section introduces the improved CenterNet network from backbone network, feature fusion module and loss function. In the fourth section, the improved model is

compared with the original CenterNet, YOLOv3 and YOLOv4 model to verify the advantages. The fifth section summarizes and analyzes the method, and points out the next improvement target.

2 CenterNet Target Detection Model

2.1 CenterNet Network Architecture

CenterNet [18] is a kind of anchor-free single-stage detection model with simple network structure and excellent detection performance. Compared with the target detection model based on anchor, CenterNet model uses the center point of the target to replace the anchor, and takes the top 100 peaks of the heatmap obtained by the feature extraction network as the center point of the target to be detected. The final center point of the target is obtained by setting the threshold. Then, according to the features of the center point, the classification and location information of the target are obtained by regression. The whole process is not based on anchor, so there is no need to set the hyper-parameter of anchor in advance. Meanwhile, the post-processing operation of NMS [19] is abandoned, which significantly reduces the amount of computation and training time of the network. The original model uses ResNet-18, DLA-34 and Hourglass-104 convolutional neural networks respectively to extract features. And it transmits the feature map to the detection module, which obtains the final result by predicting the center point and classification of the target, the size of the target, and the offset of the center point. The CenterNet model diagram is shown in Fig. 1.

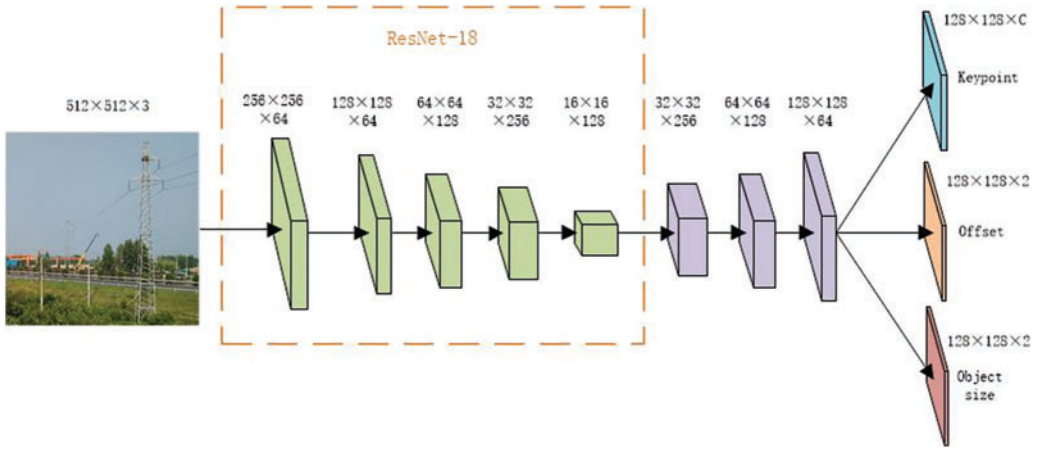

Figure 1: CenterNet network structure

2.2 Loss Function

The CenterNet detection model predicts the center point and classification, the offset of the center point and the size of the target through three independent convolution blocks. Therefore, the loss function is composed of three parts: the loss function of center point and classification, the loss function of target center offset and the loss function of target size. The loss function L_k of its center point and classification is constituted by improved focal loss [20], and its calculation is shown in formula:

$$L_k = \frac{-1}{N} \sum_{xyc} \begin{cases} (1 - \hat{Y}_{xyc})^{\alpha} \log(\hat{Y}_{xyc}), Y_{xyc} = 1 \\ (1 - \hat{Y}_{xyc})^{\beta} (\hat{Y}_{xyc})^{\alpha} \times \log(1 - \hat{Y}_{xyc}), others \end{cases} \tag{1}$$

where subscript k of L_k represents the input kth image, N is the number of key points in the image, subscript xyc is the samples of the image, Y_{xyc} is the label of the real value, $\hat{Y}_{xyc}$ is the label of the predicted value, α and β are set to 2 and 4 respectively [21].

The offset loss of the center point L_{offset} adopts L1 loss, as shown in the following formula: S_k

$$L_{offset} = \frac{1}{N}\sum_{p}\left|\hat{O}_{\tilde{p}} - \left(\frac{p}{R} - \tilde{p}\right)\right| \tag{2}$$

where p is the coordinate of the target center point in the original figure, $\frac{p}{R} - \tilde{p}$ is the offset of the predicted centers point; R is the down sampling multiple, and its value is 4.

The target size loss L_{size} also uses L1 loss, as shown in the following formula:

$$L_{size} = \frac{1}{N}\sum_{k=1}^{N}\left|\hat{S}_{pk} - S_k\right| \tag{3}$$

where S_k is the size of the original target box, and $\hat{S}_{pk}$ is the size of the target box after regression. The total loss L_{sum} is to multiply the three loss functions by the corresponding coefficients and add them, as shown in formula:

$$L_{sum} = L_k + \lambda_{size}L_{size} + \lambda_{off}L_{off} \tag{4}$$

where λ_{size} is 0.1 and λ_{off} is 1.

3 Improved Engineering Vehicle Detection Model

3.1 Overall Network Architecture

The remote detection of engineering vehicles will cause problems such as too much background information, easy occlusion, and small scale of engineering vehicles in the image, resulting in unclear target features. When detecting engineering vehicles, existing models based on deep learning fail to make full use of semantic information among feature layers, resulting in the lack of correlation among feature maps of different layers and the loss of a large number of feature information of engineering vehicles. At the same time, since the contour of the target object is not all rectangular, the bounding box of anchor-based model contains a lot of unnecessary background noise information, which is difficult to accurately express the target details, resulting in low target detection accuracy. In conclusion, this paper proposes a CenterNet based engineering vehicle detection model without anchor frame, which can better meet the detection requirements of long-distance engineering vehicles.

The overall structure of the network is shown in Fig. 2. For the input engineering vehicle image, first of all, the feature extraction is carried out using the EfficientNet-B0 network with ECA [22] module to obtain the key information of the engineering vehicle. Then, the feature maps of different scales are integrated across layers by the proposed AF-BiFPN module to enhance the feature information of engineering vehicles. Finally, the detection results are outputted by the prediction of center point and classification, target size and center point offset.

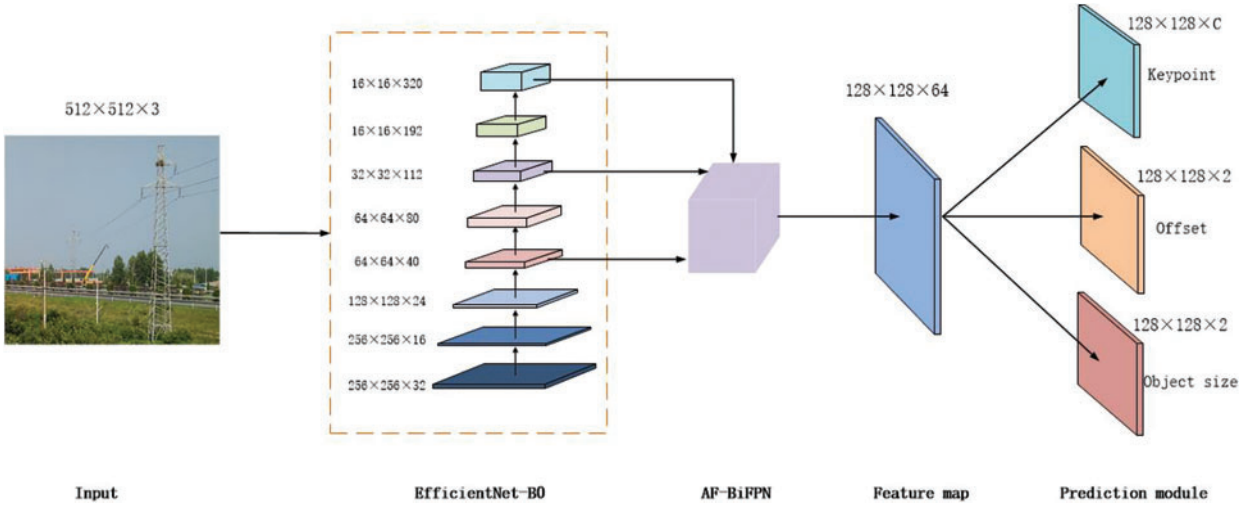

Figure 2: Overall frame diagram

3.2 EfficientNet-B0 Network with ECA

EfficientNet [23] is a model series that uses composite coefficients to weigh network depth, width and the resolution of input images, with unified control and dynamic adjustment for depth, width and resolution by defining the scaling parameter. EfficientNet has eight different versions. The EfficientNet-B0 network used in this paper has a simple structure, as shown in Tab. 1.

Table 1: EfficientNet-B0 network model

Stage	Module	Module numbers	Kernel/stide	Outputs
1	Conv0	1	(3, 3)/2	(256, 256, 32)
2	MBConv1	1	(3, 3)/1	(256, 256, 16)
3	MBConv6	2	(3, 3)/2	(128, 128, 24)
4	MBConv6	2	(5, 5)/2	(64, 64, 40)
5	MBConv6	3	(3, 3)/1	(64, 64, 80)
6	MBConv6	3	(5, 5)/2	(32, 32, 112)
7	MBConv6	4	(5, 5)/2	(16, 16, 192)
8	MBConv6	1	(3, 3)/1	(16, 16, 320)

The core structure of the network is the Mobile Inverted Bottleneck Convolution (MBConv) [24] module. This module uses the attention idea of Squeeze-and-Excitation Networks (SENet) [25] to make the model pay more attention to the more informative channel features, and suppress the unimportant channel features.

Fig. 3 shows the MBConv network structure. Firstly, the input is convolved point by point with 1∗1 convolution kernels, and the output channel dimension is changed according to the expansion ratio. Then, the k∗k convolution is used to conduct depth wise convolution operation on the extended feature map. Meanwhile, SE module is added to the channel dimension to perform weighting operation on different channels of the feature map to highlight channel features. Finally, the feature map

is convolved point by point with multiple 1∗1 convolution kernels to restore the original channel dimension.

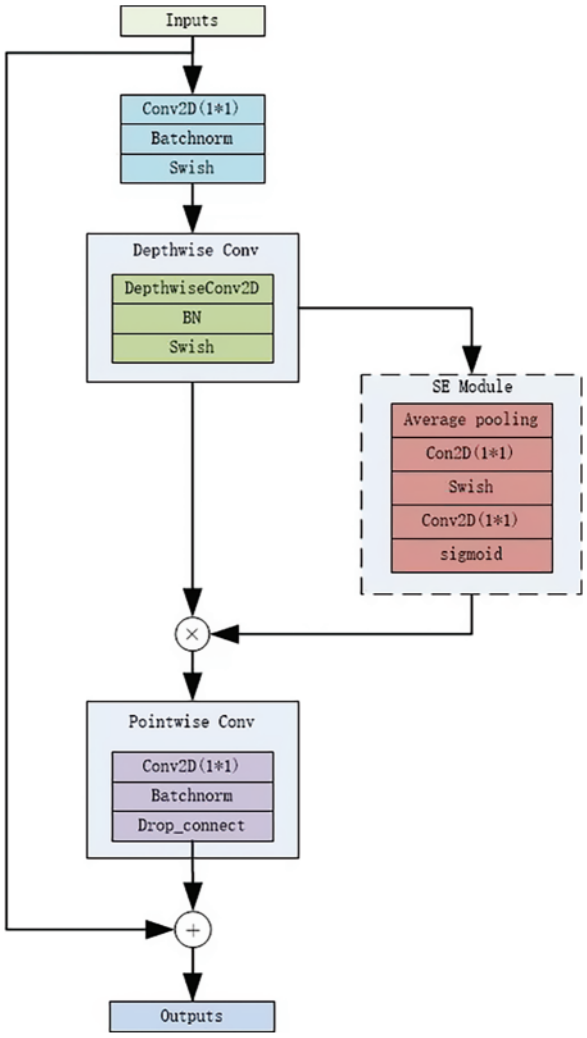

Figure 3: MBConv network structure

Since SE module compresses the channels of the input feature map, the channel and weight do not correspond directly in the dimensionality reduction process, which is not conducive to learning the relationship among channels. Therefore, this paper improves MBConv module by introducing ECA module. The compression of feature map channels is avoided by using the strategy of local cross-channel interaction without dimensionality reduction. And the coverage of local cross-channel interaction is determined by adaptive selection of one-dimensional convolution kernel, which effectively captures the information of cross-channel interaction in an extremely lightweight way. The ECA module solves the attention of the input feature channels and adaptively adjusts the importance of the channels, as shown in Fig. 4. The ECA module first carries out the operation of channel level Global Average Pooling (GAP) without reducing the dimension of the input features. Then, one-dimensional convolution is used to capture the interaction information between the current channel and its k neighborhood channels. Finally, the input features are multiplied element by element with

the weight of each channel obtained by sigmoid function to strengthen the ability to extract important features.

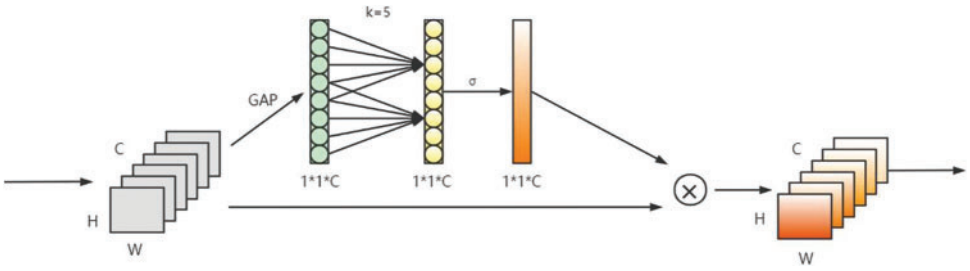

Figure 4: Structure of ECA module

ECA module uses cross-channel interaction of one-dimensional convolution of size k to replace the full connection layer, which effectively reduces the computation and complexity of the full connection layer, and then generates weights for each channel, as shown in formula:

$$\omega = \sigma(CID_k(y)) \tag{5}$$

where ω is the channel weight, σ is the sigmoid activation function, CID_k represents one-dimensional convolution of kernel k. The greater the number of channels in the input feature map, the greater the k value required for local interaction, so the k value is proportional to the number of channels C. In this paper, k value is determined by adaptive function related to channel dimension C,

$$C = \phi(k) = 2^{(\gamma * k - b)} \tag{6}$$

it can be concluded that:

$$k = \psi(\mathrm{C}) = \left| \frac{\log_2(C)}{\gamma} + \frac{b}{\gamma} \right|_{odd} \tag{7}$$

where type of $|t|_{odd}$ is the most close to the t odd, and b is set to 2, and 1 respectively.

3.3 Design of Multi-Scale Adaptive Feature Fusion Module

Due to the different shapes of engineering vehicles in the working process, and there are problems such as low target resolution and loss of details in the long-distance target detection process. It is necessary to strengthen the use of multi-scale feature fusion, which can improve the feature extraction ability of engineering vehicles and suppress the interference of noise in the target.

For convolutional neural networks, different network layers correspond to different features. Features with fine-grained, high-resolution and pixel-level can be learned in shallow networks, while high-level semantic features need to be learned in deep networks. In the detection of small targets, small size feature maps cannot provide the required resolution information, so it is necessary to combine the large ones for judgment to increase the complementarity among feature layers. Therefore, when detecting long-distance engineering vehicles, combining shallow pixel-level features with deep semantic features will have better effect.

When feature maps with different resolutions are fused, the general method is to add the different feature maps directly after adjusting their sizes to the same. However, their contribution to the output features is generally different, and different scales have inconsistency, which leads to high noise of the fused feature maps, reduces positioning accuracy and affects detection results.

The BiFPN module achieves good results in a small number of parameters by using cross-scale connections and weighted feature fusion. BiFPN network structure is shown in Fig. 5:

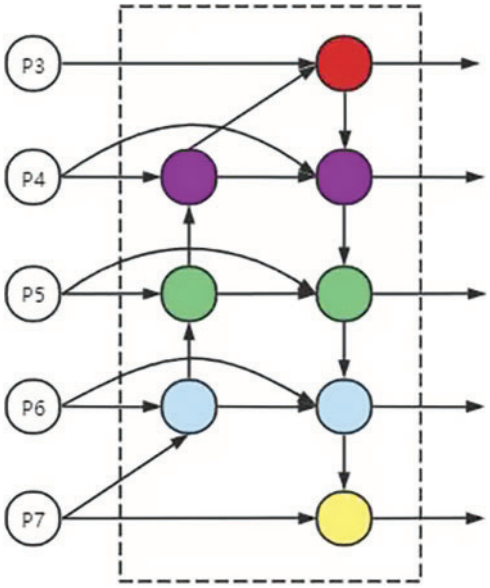

Figure 5: BiFPN network

Weighted feature fusion assigns different weights to each input feature and allows the network to learn these weights automatically. The function expression of weighted feature fusion is

$$O = \sum_i \frac{\omega_i}{\varepsilon + \sum_j \omega_j} \times I_i \tag{8}$$

where ω_i and ω_j are learnable weights, and the learning rate ε is a value far less than 1 to ensure that the denominator is not zero. It can be seen from formula (8) that the weight of feature fusion is limited between 0 and 1, and the efficiency is improved because softmax function is not used.

Formulas (9) and (10) describe the cross-scale connection and weighted feature fusion of BiFPN network at the sixth layer:

$$P_6^{td} = Conv\left(\frac{\omega_1 \cdot P_6^{in} + \omega_2 \cdot \mathrm{Re}size(P_7^{in})}{\omega_1 + \omega_2 + \varepsilon}\right) \tag{9}$$

$$P_6^{out} = Conv\left(\frac{\omega'_1 \cdot P_6^{in} + \omega'_2 \cdot P_6^{td} + \omega'_3 \cdot \mathrm{Re}size(P_5^{out})}{\omega'_1 + \omega'_2 + \omega'_3 + \varepsilon}\right) \tag{10}$$

where P^{in} is the input feature, P^{out} is the output feature, and P^{td} is the middle layer in the top-down feature fusion process.

However, when multiple feature maps are fused together, the fusion process of BiFPN module will only save the information of the lowest level feature maps, resulting in insufficient feature fusion and affecting the detection results.

In this paper, the multi-scale adaptive feature fusion module AF-BiFPN is designed, which is a combination of BiFPN module and Adaptive Spatial Feature Fusion (ASFF) module. The ASFF module adaptively adjusts the fusion ratio of different feature maps which are output by BiFPN

module. The weight coefficients of input feature maps are used as parameters for training, which makes up for the defects of BiFPN module in the process of feature map fusion. Fig. 6 shows the structure of AF-BiFPN module. The input feature maps are initially fused by BiFPN module, which obtain new feature maps L1, L2 and L3 with more feature information. Then the size of L1 and L2 is adjusted to make them the same as L3, so that L1 becomes X^{1-3}, L2 is X^{2-3} and L3 is X^{3-3}. The number of channels X^{1-3}, X^{2-3} and X^{3-3} is compressed to 16 by the 1∗1 convolution, and they are spliced along the channel direction through the concat operation. After passing through a 1∗1 convolution with three channels, softmax operation is used to make the fusion weight coefficients between 0 and 1 to obtain α^3, β^3 and γ^3 corresponding to X^{1-3}, X^{2-3} and X^{3-3}. Finally, the fusion coefficients are multiplied by X^{1-3}, X^{2-3} and X^{3-3} respectively and then added to obtain Y^3, as shown in formula:

$$Y^3 = \alpha^3 \cdot X^{1-3} + \beta^3 \cdot X^{2-3} + \gamma^3 \cdot X^{3-3} \tag{11}$$

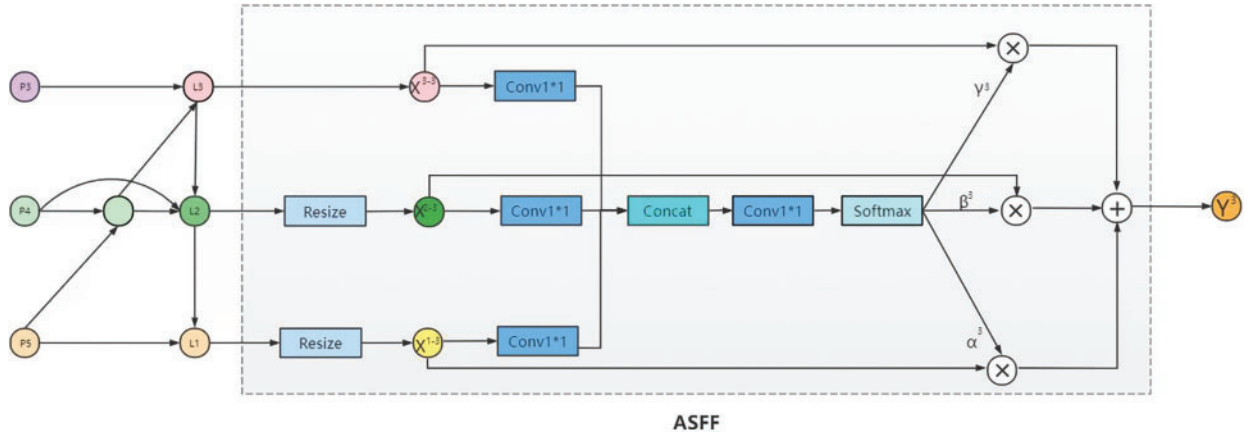

Figure 6: AF-BiFPN network structure

3.4 Design of Loss Function

The original CenterNet network model trained the center point and size of the target independently, resulting in poor positioning effect of the prediction box and easy to cause positioning deviation in the process of small target recognition. Therefore, MIOU loss is introduced in this paper to measure the positioning quality. The analysis of the predicted center point and target size represents the prediction box, and the degree of coincidence between the prediction box and the truth box is taken as the supervision item of training, so that the model can obtain a more accurate detection box. The specific calculation method is shown in Fig. 7.

The positioning quality loss can be calculated as formula:

$$L_{MIOU} = 1 - \left(IOU - \frac{m_1 + m_2 + m_3 + m_4}{d^2} \right) \tag{12}$$

where IOU represents the ratio of intersection and union between prediction and truth boxes. d represents the Euclidean distance of the diagonal of the smallest closure region that can contain both the prediction box and the truth box. $m_1 \sim m_4$ represent the Manhattan distance between the center points of the boundaries of the prediction box and the truth box respectively,

$$m_i = x_i + y_i \tag{13}$$

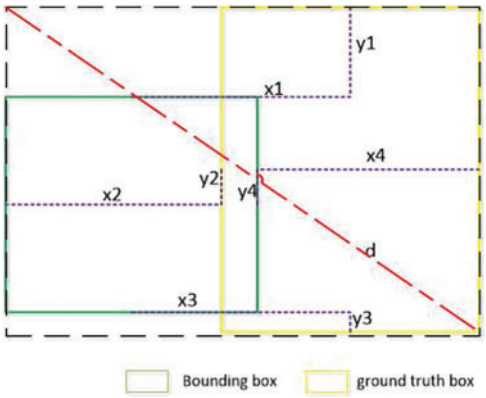

Figure 7: Calculation method of MIOU

Due to the positioning quality loss, the prediction box in the model training process gradually approaches the truth box, and the effect of the center point offset loss of the original CenterNet network is reduced. Therefore, the center point offset loss function is removed in this paper. The loss function of CenterNet model after improvement is:

$$L_{sum} = L_k + \lambda_{size} L_{size} + \lambda_{MIOU} L_{\text{MIOU}} \tag{14}$$

4 Experimental Results and Analysis

In order to verify the effectiveness of the improved CenterNet model in detecting remote engineering vehicles, the improved model was compared with other models.

4.1 Model Training Environment and Parameter Setting

The dataset required for the experiment came from UVA photographing of engineering vehicles under the high-voltage line. In order to ensure the diversity of images, they were taken under different lighting conditions and different backgrounds. A total of 1350 images including crane, bulldozer and excavator were obtained. Fig. 8 is a partial example:

Figure 8: Partial images of the dataset

In order to avoid over-fitting in the training process, the original dataset is expanded to 7200 images by flipping, rotating, cropping, brightness adjustment, adding Gaussian noise and salt and pepper noise. The expanded sample is uniformly scaled to 512×512 pixels, and the images are

randomly divided into training set, verification set and test set with the proportions of 70%, 15% and 15%. The learning rate is set to 0.001, and the batch size is 8.

The test environment was Windows10 operating system, and PyTorch deep learning framework was used to complete the environment configuration. The hardware configuration environment was Intel(R) Core i7-9700k CPU@3.6 GHz, 64.0 GB RAM, and NVIDIA GeForce GTX 1660. The software and hardware environments of the comparative experiment are the same.

4.2 Evaluation Indicators

The general evaluation standard of target detection algorithm is mAP, which is calculated by precision and recall. The calculation formula of the accuracy is as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \times 100\% \tag{15}$$

where TP is the number of prediction boxes whose IOU is greater than a certain threshold. FP indicates the number of prediction boxes whose IOU is less than a certain threshold, or the number of redundant prediction boxes with the same truth boxes. The calculation formula of recall rate is:

$$\text{Recall} = \frac{TP}{TP + FN} \times 100\% \tag{16}$$

where FN indicates the number of truth boxes that are not detected.

The Precision-Recall (P-R) curve is drawn with the precision rate as ordinate and the recall rate as horizontal coordinate. The area under the curve is the Average Precision (AP) of a certain classification:

$$AP = \int_0^1 P(R)dR \tag{17}$$

where R is recall rate and P is accuracy rate. mAP is the average AP value of all classification, and the calculation formula is as follows:

$$\text{m}AP = \frac{\sum_i^k AP_i}{k} \tag{18}$$

The detection speed is evaluated by the number of Frames Per Second (FPS).

4.3 Network Model Ablation Experiment

In order to verify the effectiveness of feature extraction network EfficientNet-B0, ECA model, BiFPN, ASFF and MIOU loss cooperative processing in the model, the ablation experiment is used to compare and analyze them. The results are shown in Tab. 2.

It can be seen from Tab. 2 that the Experiment 1 has obtained relatively good results, with mAP and FPS reaching 78.45% and 20 f/s. Experiment 2 replaces feature extraction network ResNet-18 of original detection model with lightweight convolutional neural network EfficientNet-B0. Due to its deep network layer, strong feature extraction ability, and the use of mobile inverted bottleneck convolution module, the number of parameters is greatly reduced, mAP increases by 5.82% and FPS by 10 f/s. Experiment 3 shows the effect of ECA module that the important features of the current task are acquired, and the features with minimal effect are suppressed to improve the detection effect of engineering vehicles with complex background, which improves mAP by 1.67%. Based on Experiment 3, the BiFPN module fuses shallow and deep features which increase the mAP of the model by 3.21%.

Compared with Experiment 4, Experiment 5 combined with ASFF module designes AF-BiFPN module which further enriches feature information, and improves the accuracy of construction vehicle target detection. Experiment 6 replaces the offset loss with MIOU loss by conjoint analysis of center point and target size achieves the best experimental results that the mAP reaches by 94.74%. Through the comparison and analysis of the above experiments, the detection accuracy and speed of the five improved strategies are better than the original algorithm.

Table 2: Ablation test result

Experiment	EfficientNet-B0	ECA	BiFPN	ASFF	MIOU loss	mAP/%	FPS f/s	Model size/MB
1	-	-	-	-	-	78.45	20	86
2	√	-	-	-	-	84.27	30	41
3	√	√	-	-	-	85.94	31	38
4	√	√	√	-	-	89.15	30	41
5	√	√	√	√	-	92.23	29	42
6	√	√	√	√	√	94.74	29	43

4.4 Comparison Between the Proposed Model and the Original CenterNet Detection Model

In this paper, the original CenterNet model with ResNet-18 as the backbone network and the improved CenterNet network model proposed are trained respectively. The evaluation indexes of standard COCO dataset are used. The comparison of detection results on the remote engineering vehicle data set is shown in Tab. 3.

Table 3: Comparison between the proposed model and the original detection model

Model	Crane AP/%	Excavator AP/%	Bulldozer AP/%	mAP/%	FPS f/s	Model size/MB
ResNet18-CenterNet	74	83	78	78.45	20	86
ours	91	97	95	94.74	29	43

As can be seen from the data in Tab. 3, the improved model in this paper improves the accuracy and speed of detection. Compared with the original model, the mAP of the improved model is increased by 16.29%, the FPS is increased by 9 f/s, and the model size is reduced by 43 MB. Since the engineering vehicle target photographed from a distance is small and its working shape is irregular, there are many background pixels affecting the detection result, ResNet18-Centernet is difficult to distinguish the target and background during detection. At the same time, the original model obtains heatmap by upsampling the feature map for three times to detect the target, resulting in low detection accuracy. Therefore, in this paper, the EfficientNet-B0 firstly integrated ECA module is used to replace ResNet-18 network, which ensures the feature extraction ability and reduces the number of model parameters. Then, AF-BiFPN is designed to fuse the deep and shallow features of the feature map, which enhances the accuracy of target feature extraction. Finally, in order to make the model pay attention to the location information of engineering vehicles while accurately classifying them, MIOU loss is introduced in this paper to measure the target positioning quality. The center point and the size

of the target are interpreted as detection boxes, so that the model can obtain more accurate detection effect.

The two models are used to test the same images in the data set. Fig. 9 is the comparison of detection effects. The left is the detection result of the original model ResNet18-Centernet, and the right is the detection result of the model in this paper. As can be seen from Fig. 9, although the original model can accurately detect engineering vehicles in the case of obvious targets, the model proposed in this paper has higher confidence and the prediction box is more consistent with the real location. In the case of dark color and occlusion, it is difficult to detect the target in the image because there is the small difference between the color of the engineering vehicle and the background, and the feature performance is not obvious, leading to the omission of the original model. As can be seen from the right side of Fig. 9, the proposed model can detect engineering vehicles with different environments and small targets well, which proves the robustness of the proposed model.

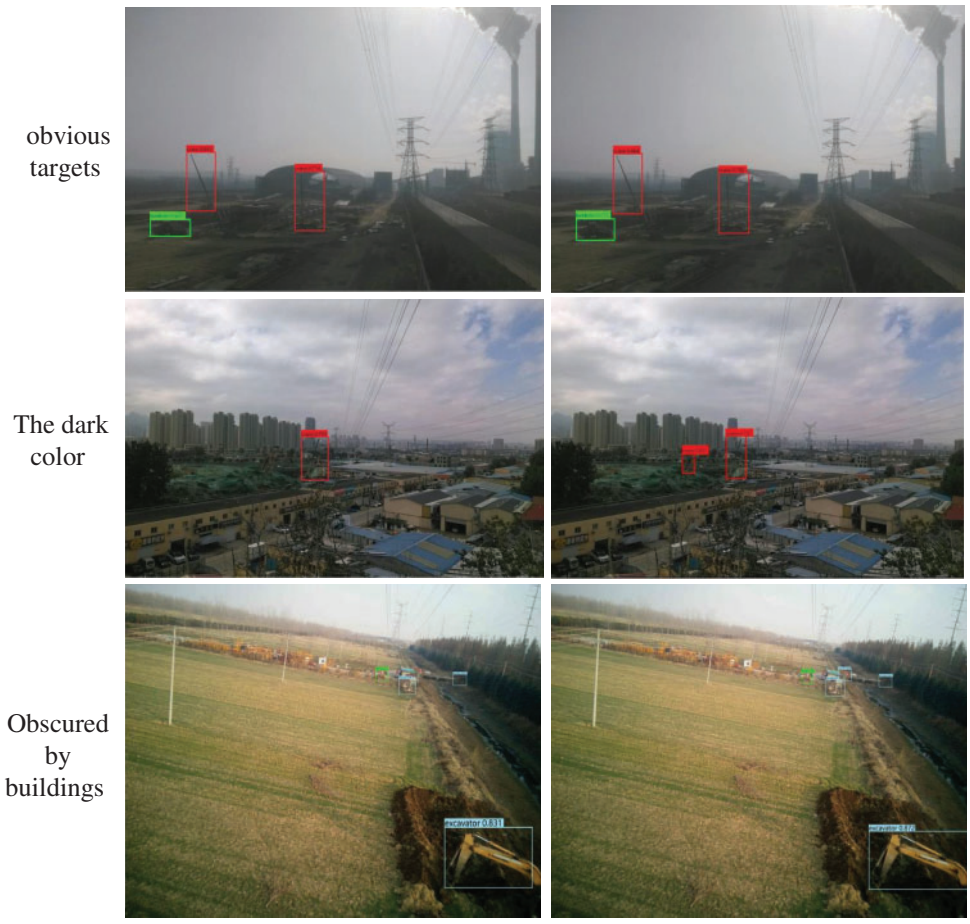

Figure 9: Comparison of detection models

4.5 Comparison Between the Proposed Model and Other Detection Models

To further verify the robustness of the proposed model, a comparative analysis is conducted with ResNet50-Centernet, YOLOv3, and YOLOv4 detection models respectively in the same experimental environment. The experimental results are listed in Tab. 4.

Table 4: Comparison between the proposed model and other detection models

Model	Crane AP/%	Excavator AP/%	Bulldozer AP/%	mAP/%	FPS f/s^{-}	Model size/MB
ResNet50-CenterNet	81	89	84	84.67	16	114
YOLOv3	73	79	78	76.84	16	216
YOLOv4	86	92	92	89.13	19	234
Ours	91	97	95	94.74	29	43

As can be seen from Tab. 4, compared with the ResNet50-Centernet detection model belonging to the anchor free series, the detection speed and accuracy of the model in this paper are improved when the model size is reduced by 71 MB. Although ResNet-50 can extract good features, it does not combine the feature information of deep and shallow layers for analysis, resulting in incomplete information of engineering vehicles in the heat maps and lower detection accuracy. At the same time, due to the large number of ResNet-50 parameters, the detection speed of the model will be reduced and the model size will be increased.

Compared with anchor-based detection models, the detection speed of the proposed model is lower than that of YOLOv3 but the detection accuracy is significantly improved, while the detection speed and accuracy of the proposed model are both improved compared with that of YOLOv4. YOLOv3 and YOLOv4 use FPN [26] and PANet [27] to fuse shallow features and deep features respectively to enrich the information of heat map. However, AF-BiFPN feature fusion network used in this paper improves the structure of PANet by removing nodes with only input edges but not fused, and adding jump connections between input and output nodes to fuse more features. By considering the detection accuracy, speed and size of the model comprehensively, the proposed model is more suitable for the detection task of engineering vehicles.

5 Concludes

In order to prevent the work of engineering vehicles from affecting the normal operation of high-voltage transmission lines, a model for detecting engineering vehicles based on CenterNet is proposed. Firstly, the EfficientNet-B0 network with smaller parameters and better feature extraction ability is used to replace ResNet-18 network in the original model. And the ECA module is used to further increase the feature extraction ability and speed of the model. Then, AF-BiFPN feature fusion network is proposed to enrich the feature information of engineering vehicles and increase the detection accuracy. Finally, MIOU loss is introduced into the loss function part to measure the positioning quality, and the center point and size of the target are combined as the prediction information, so that the model can obtain more accurate detection boxes. The results show that compared with the ResNet18-CenterNet, the proposed model improves mAP by 16.29% and detection speed by 9 f/s, achieving a good balance in detection accuracy, model size and real time, which can be better adapted

to embedded devices. As the experiments in this paper are all completed on laboratory equipment, the next step will continue to optimize the network model, improve its detection accuracy and speed, reduce the size of the model, and deploy it to embedded devices for the detection of engineering vehicles.

Acknowledgement: We would like to thank the anonymous reviewers for their valuable and helpful comments, which substantially improved this paper. At last, we also would also like to thank all of the editors for their professional advice and help.

Funding Statement: This research was funded by College Student Innovation and Entrepreneurship Training Program, Grant Number 2021055Z and S202110082031, the Special Project for Cultivating Scientific and Technological Innovation Ability of College and Middle School Students in Hebei Province, Grant Number 2021H011404.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] C. Yan, C. Wang, X. Wang, J. Du, X. Xiang *et al.,* "Based on the deep study of transmission line engineering vehicles intrusion detection," *Information Technology*, vol. 42, no. 7, pp. 28–33+38, 2018.

[2] X. Liu, Y. Zhang, S. Zhang, Y. Wang, Z. Liang *et al.,* "Detection of engineering vehicles in high-resolution monitoring images," *Frontiers of Information Technology & Electronic Engineering*, vol. 16, no. 5, pp. 346–357, 2015.

[3] Y. Zhang, K. Guo, W. Guo, J. Zhang and Y. Li, "Pedestrian crossing detection based on HOG and SVM," *Journal of Cyber Security*, vol. 3, no. 2, pp. 79–88, 2021.

[4] J. Wang, T. Zhang, Y. Cheng and N. Al-Nabhan, "Deep learning for object detection: A survey," *Computer Systems Science and Engineering*, vol. 38, no. 2, pp. 165–182, 2021.

[5] M. Hofmann, P. Tiefenbacher and G. Rigoll, "Background segmentation with feedback: The pixel-based adaptive segmenter," in *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition Workshops*, Providence, RI, USA, pp. 38–43, 2012.

[6] S. Li, J. Lin, G. Li, T. Bai, H. Wang *et al.,* "Vehicle type detection based on deep learning in traffic scene," *Procedia Computer Science*, vol. 131, pp. 564–572, 2018.

[7] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[8] M. Zhang, H. Li, G. Xia, W. Zhao, S. Ren *et al.,* "Research on the application of deep learning target detection of engineering vehicles in the patrol and inspection for military optical cable lines by UAV," in *11th Int. Symp. on Computational Intelligence and Design (ISCID)*, Hangzhou, China, vol. 1, pp. 97–101, 2018.

[9] D. Zhang, J. Hu, F. Li, X. Ding, A. K. Sangaiah *et al.,* "Small object detection via precise region-based fully convolutional networks," *Computers, Materials & Continua*, vol. 69, no. 2, pp. 1503–1517, 2021.

[10] G. H. Yu, H. H. Fan, H. Y. Zhou, T. Wu and H. J. Zhu, "Vehicle target detection method based on improved SSD model," *Journal on Artificial Intelligence*, vol. 2, no. 3, pp. 125–135, 2020.

[11] Z. Yan, Z. Jun and G. Wei, "Research on object detection of traffic scene based on deep learning," in *Proc. of the 2020 Artificial Intelligence and Complex Systems Conf.*, New York, NY, USA, pp. 133–137, 2020.

[12] D. Pu, R. Chen, X. Kong and W. Shi, "Construction vehicle detection algorithm based on deep convolution network," *Journal of Nanjing Institute of Engineering: Natural Science Edition*, vol. 19, no. 2, pp. 7, 2021.

[13] M. Zhang, T. Wang, W. Zhao, X. Chen and J. Wan, "Research on target detection of excavator in aerial photography environment based on YOLOv4," in *Int. Conf. on Robots & Intelligent System (ICRIS)*,Sanya, China, pp. 711–714, 2020.
[14] Q. Zhang, Q. Lin and L. Xiao, "Improved aerial image recognition algorithm of YOLOv5," *Changjiang Information & Communication*, vol. 34, no. 3, pp. 73–76, 2012.
[15] X. R. Zhang, X. Chen, W. Sun and X. Z. He, "Vehicle re-identification model based on optimized densenet121 with joint loss," *Computers, Materials & Continua*, vol. 67, no. 3, pp. 3933–3948, 2021.
[16] L. Zhao and M. Zhao, "Feature-enhanced refinedet: Fast detection of small objects," *Journal of Information Hiding and Privacy Protection*, vol. 3, no. 1, pp. 1–8, 2021.
[17] S. Woo, J. Park, J. Y. Lee and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. of the European Conf. on Computer Vision (ECCV)*, Munich, Germany, pp. 3–19, 2018.
[18] X. Zhou, D. Wang and P. Krähenbühl, "Objects as points," ArXiv preprint arXiv: 1904.07850, 2019.
[19] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," ArXiv preprint arXiv: 1804.02767, 2018.
[20] T. Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, "Focal loss for dense object detection," *Transactions on Pattern Analysis & Machine Intelligence*, vol. 99, pp. 2999–3007, 2017.
[21] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," *International Journal of Computer Vision*, vol. 128, no. 3, pp. 642–656, 2020.
[22] Q. Wang, B. Wu, P. Zhu, P. Li and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Conf. on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA2020.
[23] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Int. Conf. on Machine Learning*, Long Beach, CA, USA, pp. 6105–6114, 2019.
[24] M. Tan, B. Chen, R. Pang, V. Vasudevan, M. Sandler *et al.,* "Mnasnet: Platform-aware neural architecture search for mobile," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 2820–2828, 2019.
[25] J. Hu, L. Shen and G. Sun, "Squeeze-and-excitation networks," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 7132–7141, 2018.
[26] S. Liu, D. Huang and Y. Wang, "Learning spatial fusion for single-shot object detection," ArXiv preprint arXiv:1911.09516, 2019.
[27] M. Tan, R. Pang and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp. 10781–10790, 2020.