Hindawi
Discrete Dynamics in Nature and Society
Volume 2022, Article ID 8417295, 13 pages
https://doi.org/10.1155/2022/8417295

Research Article

Improved CenterNet for Accurate and Fast Fitting Object Detection

Huimin He,[1] Qionglan Na,[1] Dan Su,[1] Kai Zhao,[2] Jing Lou,[1] and Yixi Yang[3]

[1]*State Grid Jibei Information and Telecommunication Company, Beijing 100053, China*
[2]*North China Electric Power University, Department of Electronic and Communication Engineering, Baoding 071003, China*
[3]*State Grid Information and Telecommunication Branch, Beijing 100761, China*

Correspondence should be addressed to Qionglan Na; 81885883@qq.com

Received 14 January 2022; Accepted 4 May 2022; Published 30 May 2022

Academic Editor: Jorge E. Macias-Diaz

Accurate and fast detection of typical fittings is the prerequisite of condition monitoring and fault diagnosis. At present, most successful fitting detectors are anchor-based, which are challenging to meet the requirements of edge deployment. In this paper, we propose a novel anchor-free method called HRM-CenterNet. Firstly, the lightweight MobileNetV3 is introduced into CenterNet to extract multi-scale features of different layers. In addition, the lightweight receptive field enhancement module is proposed for the deep layer features to further enhance the characterization power of global features and generate more accurate heatmaps. Finally, the high-resolution feature fusion network with iterative aggregation is designed to reduce the loss of spatial semantic information in subsampling and further improve the accuracy of small and occlusion objects. Experiments are carried out on the TFITS and PASCAL VOC datasets. The results show that the size of the network is more than 60% lower than that of CenterNet. Compared with other detectors, our method achieves comparable accuracy with all accurate models at a much faster speed and meets the performance requirements of real-time detection.

1. Introduction

Transmission line fault is an important cause of power grid blackout. Inspection periodically can significantly reduce the workload of operation and maintenance personnel on pole inspection, which is an essential means to ensure the safe operation of the power system [1]. Fittings are iron or copper metal accessories widely used in transmission lines, mainly used to support, fix, and connect bare conductors, conductors, and insulators [2]. Because the fittings run outdoors all year round, it is easy to produce corrosion, deformation, damage, and other phenomena. Therefore, the realization of high-precision automatic detection of fittings can predict their faults in advance, which is of great significance in ensuring the safe operation of the power grid [3]. In recent years, the power system has promoted unmanned aerial vehicle transmission line inspection for its high security and efficiency. It can also be combined with object detection technology to realize intelligent processing [4]. Besides, the use of computer vision and image processing technology for aerial images and automatic video processing can realize the automatic fault location of transmission line fittings, which significantly improves the efficiency of power maintenance.

With the successful application of deep convolutional neural network (DCNN) [5], object detection performance has improved significantly. At present, most of the existing fitting detectors are anchor-based [6]. Researchers are committed to improving the accuracy and efficiency of anchor-based methods. Although the stability of these algorithms has been improved, it has high requirements for hardware computing resources due to too many model parameters. In addition, the detection speed is slow, which cannot meet the needs of real-time detection. At the same time, the large size of the model also makes it unable to apply to the operating platform with relatively limited hardware resources.

To solve these problems, we put forward anchor-free detectors to improve the flexibility of the fitting detection. Anchor-free methods do not depend on the preset anchors

but adapt different fitting objects through regression. However, these detectors represented by CenterNet [7] also face some problems. First, they cannot achieve a good trade-off between efficiency and accuracy in practical applications. The accuracy of the faster network is ordinary, while the efficiency of the network with higher accuracy cannot meet real-time detection requirements. Secondly, the anchor-free detectors adopt a simple design and no feature fusion operation, which leads to the problem of mutual interference between objects of different scales widely, especially in the detectors using an hourglass network as the backbone. Finally, these detectors only generate a single-scale feature map. The insufficient feature extraction leads to low accuracy of small fittings (such as hanging board, u-type hanging ring) and those with occlusion (such as shockproof hammer and yoke plate). In addition, the low-resolution feature map also causes objects with huge scale gaps to be mixed into one feature map, so the accuracy is ordinary for complex aerial fitting images [8].

The performance of fitting object detection can be further improved by designing a novel backbone network and integrating the feature fusion method. Compared with anchor-free methods, anchor-based detection accuracy is better than the former, mainly due to the feature fusion network (FPN) [9]. FPN integrates the features of various scales to reduce feature loss in the upsampling process. Inspired by this idea, we propose a novel anchor-free method by introducing the lightweight backbone network and designing the feature fusion structure to CenterNet. Our method is called high-resolution MobileNet-based CenterNet (HRM-CenterNet), which can achieve better detection accuracy with high efficiency. Figure 1 shows the diagram of our method.

The main contributions of this paper are listed as follows. Firstly, it proposes the lightweight MobileNetV3 as the backbone to extract convolution features of different layers, reducing the size of the model and detection time-consuming while maintaining high accuracy. Secondly, it improves the lightweight feature receptive field enhancement module, which can expand the receptive field of high-layer features, enhance the feature expression ability of high-layer semantics, and strengthen its context information. Thirdly, it designs the high-resolution feature fusion network based on iterative aggregation. It can reduce the spatial features lost due to continuous down-sampling, maximizing the use of feature information extracted from the backbone network and generating heat maps more accurately.

We note that a shorter conference version of this manuscript appeared in Zhao et al. [10]. Our initial conference paper did not address the problem of HRM-CenterNet's effectiveness on mutual interference between fittings of different scales. This manuscript proposes the lightweight receptive field enhancement module to address this issue and provides additional performance comparison analysis with other state-of-the-art detectors. Our method effectively improves the fitting detection performance of CenterNet and achieves the best trade-off between accuracy and efficiency.

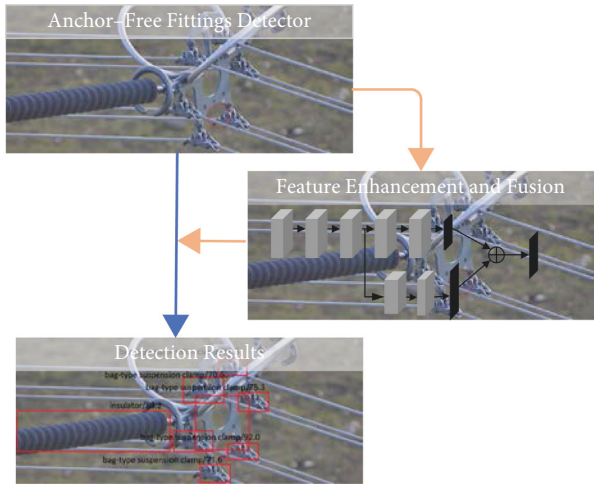


Figure 1: Diagram of our method.

2. Related Work

Object detection has always been the focus of research in computer vision, and it is also one of the difficulties in this field. Traditional detection technology uses artificially designed features, such as edge detection symbols, AdaBoost, or combines shallow features such as color, shape, and texture for recognition [11, 12]. These methods are rigid, with low accuracy and poor generalization ability. Thanks to the powerful feature extraction function of convolutional neural network(CNN), deep learning has gradually replaced traditional machine vision as the mainstream methods in image classification, object detection, and semantic segmentation [13–16]. Object detectors based on CNN can extract image features more effectively and perform end-to-end training. The existing methods can be divided into anchor-based detectors and anchor-free detectors.

2.1. Anchor-Based Detectors. The anchor-based detectors inherit the traditional idea of sliding window and region proposal strategy. Most detectors take multiple width-to-height ratio anchors as reference points for object positioning and then calculate the intersection over union (IOU) of bounding boxes between ground truth and prediction to select the most accurate one [17]. Whether there is a region proposal process can be divided into single-stage detectors and two-stage detectors.

2.1.1. Single Stage. The single-stage object detection algorithm arranges the possible bounding boxes on the image in a complex way and classifies them using the sliding window method without generating proposal regions. Moreover, feature maps are directly generated through the convolution network to predict category probability and position coordinates. Two typical single-stage detectors are YOLO [18] and SSD [19].

YOLO uses a single network for detection, which improves detection speed and has strong generalization ability. YOLOv2 [20] introduces the anchor box idea of Fast R-CNN

and uses the k-means clustering method to generate anchors. YOLOv3 [21] uses multi-label classification and cross-scale prediction methods, and it introduces Darknet-53 [22] as the feature extraction network to improve the detection accuracy of the model. YOLOv4 [23] uses CSPDarknet-53 as the backbone network, and the neck adopts SPP additional blocks and PANet [24] path aggregation blocks. It also achieves an excellent trade-off between detection speed and accuracy. SSD combines the regression idea of YOLO with the anchor mechanism of Faster R-CNN [25] to extract feature maps with different resolutions for detection [26].

2.1.2. Two Stage. In the two-stage detectors, the image is passed through a pretrained CNN to extract high-level features firstly. The region proposal network (RPN) is applied to attain two outputs: the probability that the region has an object and the coordinates of the bounding box. The RPN is trained to efficiently extract a predefined number ($k = 2000$) regions from images. R-CNN [8], one of the first successful detectors, selects the region of interest and then sends it to the subsequent convolutional neural networks (CNNs) for classification. Fast R-CNN [27] cuts the image features of different scales to reduce the amount of calculation and then sends the features to the classification and regression layer to calculate the final results. However, these two methods have the disadvantage of relying on manual feature extraction to obtain the region of interest. Fast R-CNN proposes the RPN, which replaces the traditional region of interest generation method. Mask R-CNN [28] uses ROIAlign instead of ROIPooling and adds mask branch and corresponding loss based on Fast R-CNN to achieve better classification performance. Cascade R-CNN [29] designs the cascade detection network and detects based on different IOU thresholds.

The anchor-based detectors improve the accuracy but bring many disadvantages, such as too many super-parameters and an imbalance of positive and negative samples. Besides, it needs much practical experience to design abundant anchors and assign them to specific objects. When the anchors use interest over union (IOU) as the evaluation criterion to determine the object, different IOU thresholds will significantly fluctuate the algorithm performance. The tuning processes of anchors are usually time-consuming and laborious. For complex transmission line inspection scenarios with diversified objects, the applicability of the anchor-based method is limited.

2.2. Anchor-Free Detectors. Anchor-free detectors use key point estimation for object detection. They do not need NMS postprocessing operation and preset anchors [30], which provide a new idea for achieving high-precision real-time fitting detection.

CornerNet [31] applies key points to object detection for the first time. It transforms the object position detection into detecting key points in the upper left corner and lower right corner of the object bounding box. The introduction of anchor-free approaches greatly simplifies the output, but at the same time, the accuracy is better than the one-stage anchor-based detectors. The ExtremeNet [32] detects four extreme points and one center point through a standard key point estimation network and groups the key points using the geometric relationship. A group of extreme points corresponds to a detection result. The CenterNet (key point triplets for object detection) [33] constructs triple key points based on CornerNet. Each object is represented by a center key point and a diagonal point. The CenterNet (objects as points) simplifies the problem of detecting paired key points into the estimation problem of center key points and uses the regression method to obtain objects' category, width, and height. The model is more concise and speeds up the overall speed of the algorithm [34]. CenterNet requires only one central point to locate the objects, which is one of the best anchor-free networks. FCOS [35] obtains the detection result of the input image by regressing the distance between the pixel and the left, top, right, and bottom edges of the object bounding box through the regression operation of the feature map pixel level. The pixel-by-pixel operation also brings the problem of slow detection speed.

In general, object detection by center point estimation can effectively adapt to the fitting object with variable size. In addition, it can meet the needs of real-time detection efficiency. However, the accuracy in complex fitting scenes is ordinary. To deal with that, we introduce feature receptive field enhancement and high-resolution methods to improve fitting detection accuracy.

2.3. Fitting Object Detectors. Driven by the wave of deep learning, transferring the detectors that perform well in general object detection to power field detection has become a research hotspot of power system recognition and detection. In addition, some of the early research used nondeep learning algorithms, which mainly use the color, gradient, and contour of the fitting object to recognize the metal tools. [36] uses saliency detection based on color and gradient features to locate insulators and then uses adaptive morphology to detect self-exploding defects. [37] applies F-PISA clustering to locate insulators based on color and structural characteristics and established a color model to identify damaged areas. [38] proposes GrabCut segmentation algorithm extract insulator contour and calculated the changes of gap and overhang of insulator string umbrella cap based on the convex defects of insulator contour, to quantitatively analyze the ice coating condition of insulators. [39] adopts multi-layer perceptron to detect the insulators and shock hammers based on the detected object's location correlation feature and local contour.

At present, the transmission line fitting detectors based on deep learning algorithm are mainly anchor-based method. Single-stage models such as SSD and Yolo series and two-stage model such as Faster R-CNN are classic algorithms. [40] applies the Faster R-CNN to the detection of grading ring and shockproof hammer. Aiming at the insulator detection task in aerial images, [41] proposes an object decomposition and aggregation algorithm based on YOLOv3. The above algorithms have achieved good results in fittings with a large proportion of grading ring, insulator,

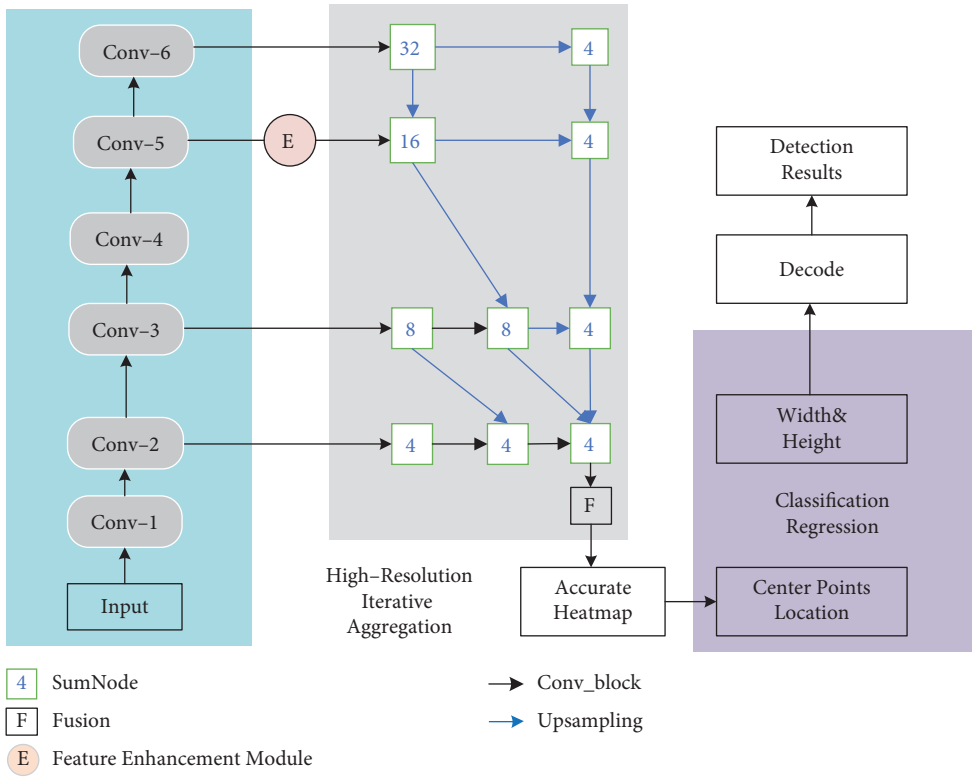


Figure 2: Structure of HRM-CenterNet.

and shockproof hammer in the inspection image and single shape. However, the complex environment of the transmission line leads to the changeable background of aerial images and multi-scale fitting objects. Hence, the detection effect of directly using the above algorithms is poor. [42] puts forward the description method of occlusion relationship between fitting objects and designs the occlusion relationship module to improve the accuracy of occlusion fittings. [43] improves the performance of SSD on dense occluded fittings by expanding the sample of metalware and introducing effective repulsion loss for dense object detection. [44] proposes a typical Faster R-CNN fitting detector combining KullbackLeibler divergence and shape constraints, which solves the problem of inaccurate object bounding box regression in fitting detection to a certain extent.

Recently, anchor-free detectors have been widely used in various industrial scenarios, including the electric power system. [45] adopts the CenterNet combined with structured positioning to realize the accurate identification and positioning of different substation equipment and components. It shows that the anchor-free method effectively improves the detection accuracy of power images and provides a new method for intelligent detection of fitting images. [46] uses DLANet backbone network, deep layer aggregation, sequence and exception module, and deformable convolution to design an efficient deep feature extraction network DLA-SE on the CenterNet to achieve the real-time detection of three common inspection faults: insulator self-explosion, shockproof hammer falling off, and bird's nest.

3. HRM-CenterNet Approach

The proposed HRM-CenterNet is an anchor-free detection method. We redesign MobileNetV3 [47] as the backbone of CenterNet. Built on M-CenterNet, the lightweight receptive field enhancement module and high-resolution feature fusion network are proposed to extract features better. Figure 2 shows the structure of HRM-CenterNet.

The key idea of the CenterNet is to predict the center point of the object through the key point heatmap and then regress the object's size, 3D position, and pose attributes from other feature maps corresponding to the key points. Compared with other anchor-free detectors, CenterNet locates the objects through only one center point with fewer parameters and a faster detection speed.

We assume that the input image is $I \in R^{W\times H\times 3}$, and W and H are the width and height of the fitting image, respectively. After passing through the backbone network, the key point heatmap $\widehat{Y} \in [0,1]^{W/R\times H/R\times c}$ is generated, R is the scale of the heatmap size, and c is the number of fittings. Then, three prediction branches are generated from the heatmap: the key point prediction branch is used to detect the key points of the heatmap and the object center point. The object size prediction branch generates a prediction bounding box based on the center point to detect the width

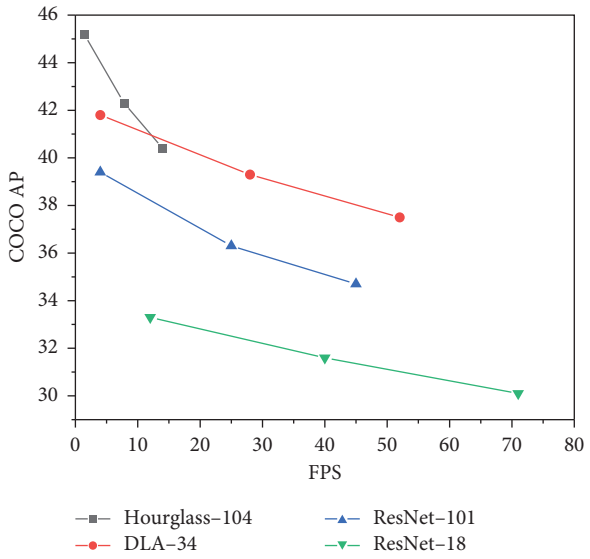


Figure 3: Performance comparison of origin backbone.

and height. The local offset prediction branch regresses the offset of the key point and the center point to detect the object accurately.

3.1. M-CenterNet. CenterNet adopts the design method of the backbone and downsampling path as a whole, which has robust scalability. The author uses four different backbone networks, but the performance has apparent differences. As shown in Figure 3, the high-speed network has general accuracy, while the high-precision network speed cannot meet the real-time requirements.

The transmission line inspection platform usually adopts a small-size edge-end device to ensure flexibility and portability. Such systems are relatively scarce of computing and storage resources, and they are more sensitive to the size and speed of the model. Choosing a lightweight network is the only way to achieve real-time detection. However, the original four backbone networks cannot effectively balance speed and accuracy, so it is necessary to use a novel backbone network that takes both into account.

The MobileNetV3 is a lightweight network focusing on mobile terminals and embedded devices. It improves the existing deep separable convolution (DSC) [48] inverted residuals and proposes a squeeze-and-excitation block (SE block). DSC decomposes the standard convolution operation into depthwise convolution and pointwise convolution, so the parameters are greatly reduced. The inverted residuals first expand the channels of the input feature map, then use DSC for downsampling to reduce the size of the feature map, and finally build the channels to enhance the expression ability of the model. SE block uses global pooling to generate channel statistical information for compression operation, compresses the global spatial information into a channel descriptor, passes through two fully connected layers, and finally uses sigmoid activation function for activation operation. Because of the advantages of MobileNetV3 in lightweight and speed, it is proposed as the backbone network of our model to extract useful features from fitting images. In this paper, MobileNetV3-large is selected as the feature extraction network for fitting detection.

In order to obtain more effective feature maps for high-precision detection, we have made further improvements to MobileNetV3. We remove the average pooling layer and three 1×1 convolution layers of the last bottleneck layer of the MobileNetV3 and then add three upsampling transposed convolution layers to restore the semantic and location image information. The feature maps generated by transposed convolution are sent to the three sub-networks of CenterNet for key point, offset, and size prediction. The network is called M-CenterNet, and the structure is shown in Figure 4.

3.2. Lightweight Receptive Field Enhancement Module. In the detection task of fittings, there are different scales and different kinds, and the same kind of fittings also has different scales. Because of the different distances and angles between aerial images and cameras, even the same fittings have large differences. It is difficult to solve mutual interference between fittings of different scales only by relying on single feature extraction and simple network design. In addition, in the top-down decoding and fusion feature process of M-CenterNet, the deep abstract features representing the semantic meaning of significant objects will be gradually diluted by the shallow representation information, thus losing the guiding spatial information.

Based on the idea of RFBNet [49], we propose the lightweight receptive field enhancement module (LRFEM) captures different levels of context information by designing multi-branch and multi-scale perforated convolution to enhance the robustness of global features.

The RFBNet simulates the relationship between the size and eccentricity of receptive fields (RFs) in the human visual system, enhances the feature extraction ability and robustness, and achieves high detection accuracy while considering the efficiency. It draws on the idea of the inception algorithm and introduces three dilated convolution layers, which effectively increases the receptive field of the network. The structure is shown in Figure 5.

The dilated convolution adds the dilation rate to each conventional convolution layer. The dilation rate determines the distance between pixels when the convolution kernel processes data, expands the kernel to the specified scale, and fills the unoccupied pixel area in the original kernel with 0. Therefore, the receptive field of dilated convolution will be improved compared with conventional convolution without increasing the amount of calculation. The calculation formula of dilated convolution receptive field is shown in formula (1): K is the receptive field of dilated convolution, rate is the dilation rate, and k is the size of the convolution kernel.

$$K = (\text{rate} - 1) \times (k - 1) + k. \tag{1}$$

In order to reduce the amount of calculation, RFB-s is improved based on RFB. The structure is shown in Figure 6(a). On the one hand, 3×3 convolution layer is used

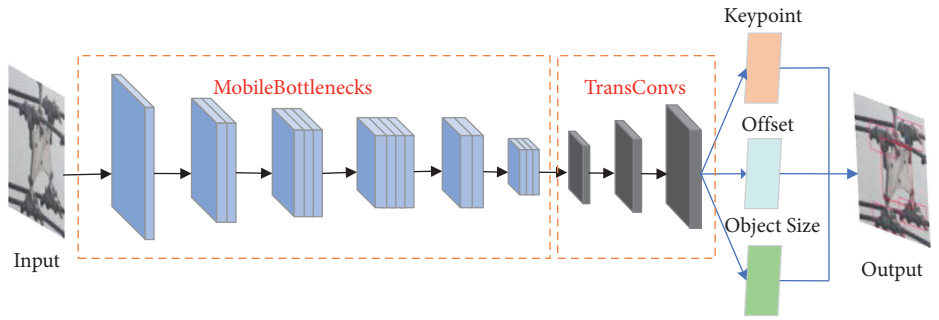


Figure 4: Structure of M-CenterNet.

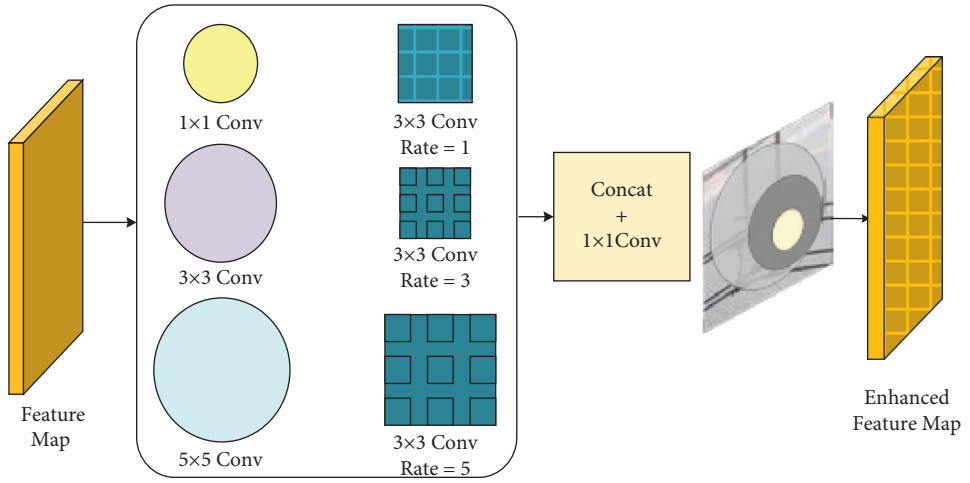


Figure 5: Structure of RFB.

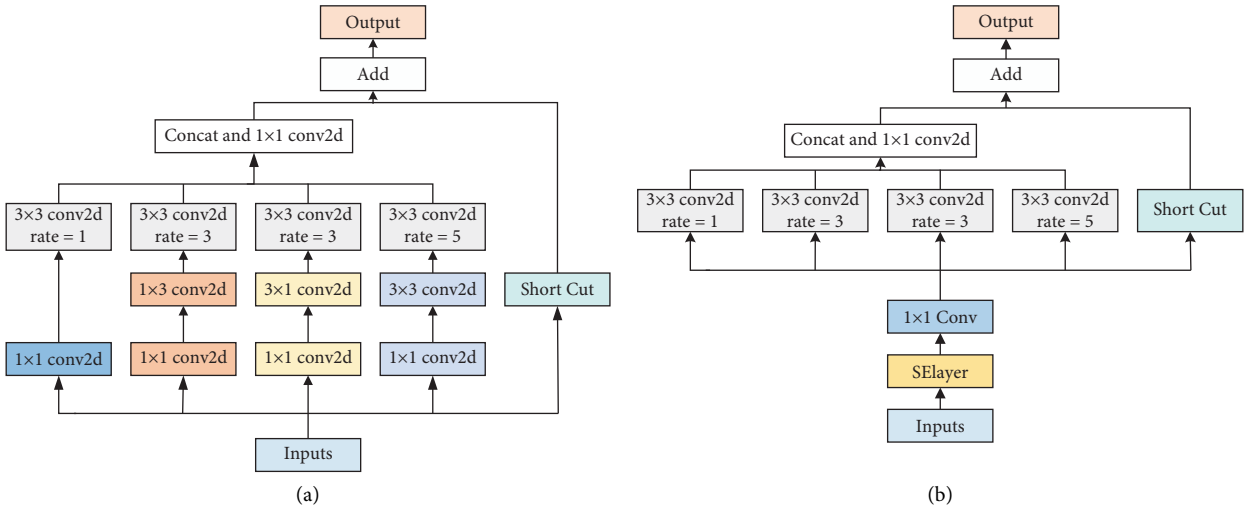


Figure 6: Structure comparison of RFB-s and our improved lightweight RFB module. (a) Structure of RFB-s. (b) Structure of improved LRFBM.

to replace 5×5 convolution layer, and on the other hand, 1×3 and 3×1 convolution layers are used to replace 3×3 convolution layer. Considering that MobileNetV3 has too many convolution layers, we appropriately delete some convolution layers in the RFB-s to reduce the calculation consumption of the network and avoid the feature map being too small. The structure is shown in Figure 6(b). The input feature map first passes through the squeeze-and-excitation layer (SE layer) to select the channel and then goes through the 1×1 convolution layer to reduce the dimension and connect four 3×3 dilated convolution layers in parallel. After passing through the 1×1 convolution layer, they are

integrated with the previous convolution layer for the second time. The results show that after the module deletes these convolution layers, the detection accuracy of the model is equivalent to that before, and the computational efficiency is significantly improved. In our work, to not excessively increase the amount of calculation, we mainly apply the LRFEM to the 16 times downsampling feature map.

3.3. High-Resolution Feature Fusion Method. The network usually connects the feature extraction module composed of convolution operation in series and generates a single-scale feature map among the anchor-free detectors. Although the low-resolution feature map output after multiple downsampling contains rich high-level semantic information, it also loses much spatial information.

For example, when ResNet18 is used as the backbone of CenterNet, the ResNet18 performs 32 times downsampling on the input image, resulting in two adjacent fittings in the original image which will become the same pixel or even disappearing. Therefore, it is difficult to achieve high-precision detection for small-size fittings and fittings with mutual occlusion.

In addition, the author only uses the largest feature map to generate the heatmap in the CenterNet, but the loss of image features leads to inaccurate generation. In order to make full use of the feature map generated after convolution operation, we propose a high-resolution feature fusion network based on iterative aggregation by referring to the high-resolution representation network.

HRNet [50] maintains high-resolution representation and gradually adds parallel subnets in feature extraction, using the feature maps extracted from the subnets for multi-scale feature fusion. It effectively utilizes feature maps with different resolutions and has a better prediction effect on heatmaps. Figure 7 shows the feature fusion method of the original HRNet. Firstly, the low-resolution feature map is sampled to the same scale as the high resolution; then, the four feature maps are concatenated. Compared with other networks characterized by aggregating layers by upsampling from low to high, HRNet has higher parameter efficiency and lighter weight.

Although the original HRNet utilized features with different resolutions, the fusion method was too simple to fully use features with different resolutions. Therefore, we designed the feature fusion method of iterative aggregation, as shown in Figure 8.

It integrates low-resolution and high-resolution features through iterative aggregation, maximizing the use of feature information extracted from the backbone network and generating heat maps more accurately. The formula of iterative aggregation is equation (2). $x_1, \ldots, x_n$ represents the aggregation node and is the input of the aggregation node.

$$I(x_1, \ldots, x_n) = \begin{cases} x_1, & \text{if } n = 1, \\ I(N(x_1, x_2), \ldots, x_n), & \text{otherwise.} \end{cases} \quad (2)$$

We remove the transpose convolution of the M-CenterNet and then select four feature maps obtained from 4

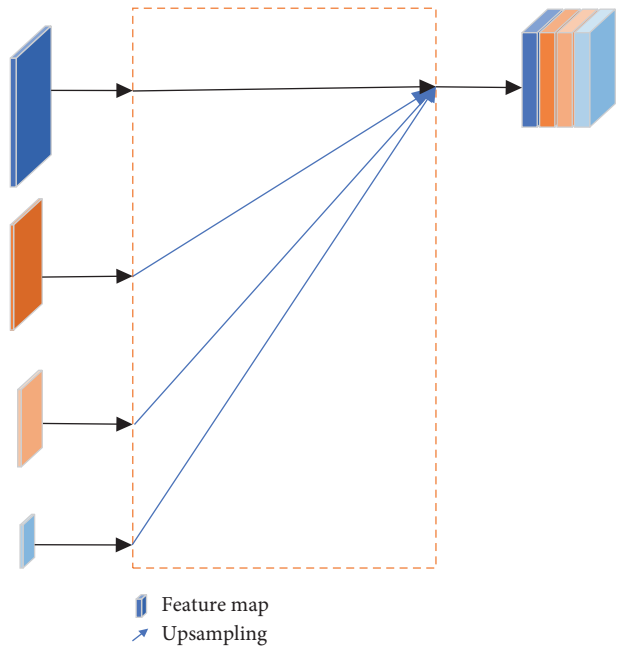


Figure 7: Feature fusion of original HRNet.

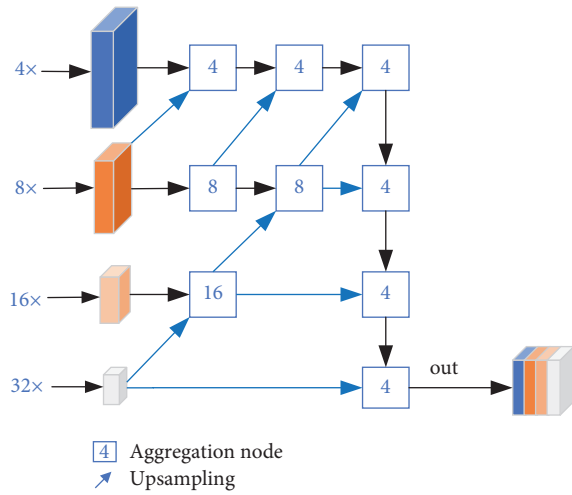


Figure 8: High-resolution feature fusion method.

times of downsampling feature map C2, 8 times of downsampling feature map C3, 16 times of downsampling feature map C5, and 32 times of downsampling feature map C6 for high-resolution feature fusion. The structure of HRM-CenterNet is shown in Figure 2. It increases the depth of the network and improves the learning ability of difficult samples. Due to the different sizes of the input feature maps of each aggregation node, we sample the low-resolution feature maps to the same size as the high-resolution feature maps through transpose convolution.

4. Experiments

In this section, we evaluate the performance of the HRM-CenterNet on TFITS and PASCAL VOC. The experiments

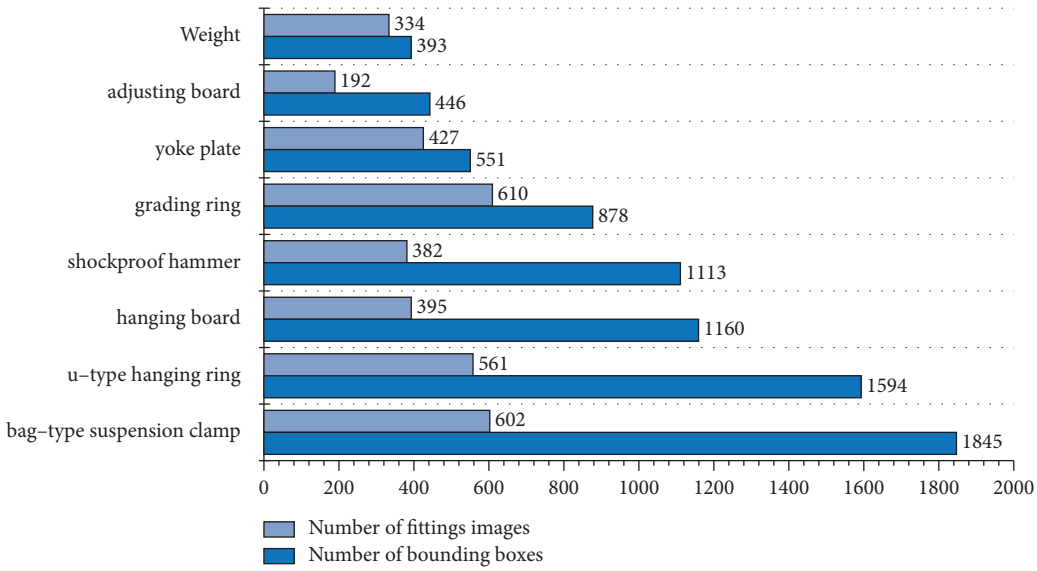


Figure 9: Quantity distribution of images and bounding boxes.

Figure 10: Aerial image of complex transmission lines in TFITS.

are implemented in the PyTorch on a machine with GeForce GTX1080Ti GPUs, CUDA 9.0, and cuDNN v7.

4.1. Implementation Details

4.1.1. Experiments on TFITS. Referring to the construction method of the PASCAL VOC dataset, we select the aerial images of UAV according to the requirements of the standard for shooting position, exposure degree, and focusing accuracy, and then construct the professional Typical Fittings Dataset (TFITS). The TFITS Dataset includes 8 types of fitting objects, consisting of 3503 images and 7980 boxes. The specific number of images and bounding boxes is shown in Figure 9. Some examples are shown in Figure 10.

We adopt the mini-batch stochastic gradient descent (SGD) momentum method to train HRM-CenterNet. The input images are uniformly scaled to 512×512. The initial learning rate is set to 1.25×10^{-4}, and the batch size is 8. The same learning rate is adopted for all layers. In the training process, when the detection accuracy of the verification set is no longer improved, the learning rate is reduced to 10% of the current learning rate by cosine annealing until the accuracy is no longer improved by adjusting the learning rate. It is trained for 140 epochs, and flip augmentation is used in testing.

4.1.2. Experiments on PASCAL VOC. The PASCAL VOC dataset contains 20 categories, which are quite different in object size, direction, posture, brightness, and occlusion position. We focus on the categories prone to false and missing detection, such as chairs, dining tables, and potted plants.

We experiment HRM-CenterNet in a small training resolution. The input images are 384×384, while all other hyper-parameters in the loss function are the same as the TFITS experiments. The network is trained on two GPUs with the batch size of 32. Adam optimizer is used with an initial learning rate of 1.25×10^{-4}. It is trained for 160 epochs, and flip augmentation is also used in testing.

4.2. Evaluation Index. In order to analyze the performance of the fitting detectors, we need to consider their accuracy and efficiency. At the same time, considering the limited

Table 1: Results on TFITS test set.

Methods	Backbone	mAP/%	FPR	FNR	Size/Mb	FPS
Single-stage detectors						
SSD [19]	ResNet101	74.1	28.4	23.5	90.6	13.1
YOLOv3 [21]	DarkNet53	75.0	26.5	20.5	59.6	20.0
YOLOv4 [23]	DarkNet53	78.6	23.1	19.2	30.0	25.0
RetinaNet [51]	ResNet101	79.1	22.3	19.3	65.3	10.2
Two-stage detectors						
Fast R-CNN [27]	ResNet101	77.1	22.8	18.4	98.0	8.8
Faster R-CNN [25]	ResNet101	80.2	20.0	15.3	122.0	7.0
Anchor-free detectors						
ExtremeNet [32]	Hourglass104	80.9	19.6	14.7	150.6	16.8
CornerNet [31]	Hourglass104	77.9	23.2	19.8	160.0	9.2
EfficientDet [52]	EffcientNetB0	74.0	26.1	20.8	80.6	47.0
CenterNet						
CenterNet-HG	Hourglass104	78.1	23.8	19.4	220.3	14.7
CenterNet-Res18	ResNet18	69.9	30.2	25.9	60.3	55.0
CenterNet-Res50	ResNet50	74.6	27.6	23.5	128.0	48.5
CenterNet-Res101	ResNet101	76.0	25.1	20.6	180.0	22.5
CenterNet-DLA34	DLA-34	76.4	24.0	20.1	77.0	27.2
M-CenterNet	MobileNetV3	75.9	25.3	20.9	18.5	43.2
HRM-CenterNet	MobileNetV3	80.3	19.8	13.6	24.6	32.5

HRM-CenterNet							
HRFF	LRFEM	Backbone	mAP/%	FPR	FNR	Size/Mb	FPS
		MobileNetV3	75.9	25.3	20.9	18.5	43.2
√		MobileNetV3	79.0	22.1	16.6	24.0	33.0
√	√	MobileNetV3	80.3	19.8	13.6	24.6	32.5

hardware devices of the edge, we also need to compare the computer memory occupied by the model. Therefore, we quantitatively evaluate the comprehensive performance of the detectors from the three dimensions: detection accuracy, efficiency, and model size.

We adopt the average accuracy, false detection rate, and missed detection rate as the evaluation indexes of the detection accuracy. Model parameters and size can better reflect the occupation of hardware computing and storage resources. The calculation methods of false and missed detection rates are shown in equations (3) and (4) The FP indicates that a nonexistent object is predicted. FN indicates that no existing object is predicted. TP indicates that the existing object is correctly predicted, and TN indicates that the algorithm correctly predicts the background.

$$\text{false detection rate} = \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}. \tag{3}$$

$$\text{missed detection rate} = \text{FNR} = \frac{\text{FN}}{\text{TP} + \text{FN}}. \tag{4}$$

4.3. Structure Experiments on TFITS Validation Set. We horizontally compare the performance of the detectors before and after improvement in this section. Table 1 shows the performance of different detectors. Table 2 shows the average accuracy of CenterNet-HG, M-CenterNet, and HRM-CenterNet.

It can be seen that our methods have improved the detection accuracy of each category. Among all kinds of fittings, the AP of the hanging board, u-type hanging ring, yoke plate, shockproof hammer, and adjusting board increases most obviously. Besides, the plate fittings such as yoke plate, adjusting board, and hanging board have large shape differences due to different shooting angles and are easily confused with tower materials. After adding the multiscale feature receptive field enhancement module, the network can significantly improve the accuracy of such fittings. In addition, some small fittings such as the hanging board, u-type hanging ring, and shockproof hammer are densely distributed in the image, resulting in a large number of occlusions among the fittings. The high-resolution feature fusion method can significantly improve the accuracy of such fittings.

We can see that the false detection rate and missed detection rate of our HRM-CenterNet are 19.8% and 13.6% in Table 1, which are better than before. Besides, compared with baseline M-CenterNet, the detection accuracy is improved by 4.4%. The false detection rate and the missed detection rate are lower by 5.5% and 7.3%, which shows that our method can improve the feature extraction ability to distinguish the fitting object and background better. Compared with the CenterNet-HG, which has the highest accuracy in the four original backbone networks, our method improves the accuracy by 2% and the efficiency by more than twice, meeting the requirements of real-time detection.

4.4. Ablation Experiments on TFITS Validation Set. The baseline for all experiments in this section is the basic M-CenterNet, which only introduces the MobileNetV3 as the backbone network. We explore the effectiveness of

Table 2: Precision comparison on TFITS.

Category	CenterNet-HG	M-CenterNet	HRM-CenterNet
Grading ring	88.3	86.3	89.3
Shockproof hammer	85.8	83.7	89.6
Bag-type suspension clamp	89.6	89.0	91.8
Yoke plate	70.8	68.4	73.9
u-type hanging ring	60.7	59.5	64.2
Hanging board	59.6	56.1	60.7
Adjusting board	78.3	76.2	80.4
Weight	92.0	88.6	92.2
mAP	78.1	75.9	80.3

introducing lightweight MobileNetV3, the lightweight receptive field enhancement module (LRFEM), and the high-resolution feature fusion (HRFF) network. The experiments are carried out on the TFITS Dataset.

Firstly, we verify the influence of different backbones on the performance of the detectors and the superiority of adopting MobileNetV3. The results in Table 1 show that the detection speed of the M-CenterNet network can reach 43.2FPS, far exceeding the CenterNet with Hourglass104 and DLA-34 as the backbone network. In addition, compared with the CenterNet with Hourglass104, DLA-34, ResNet101, ResNet50, and ResNet18 as the backbone, the size of M-CenterNet is reduced by 91.6%, 75.9%, 69.3%, 85.5%, and 89.7%, respectively, greatly reducing the occupation of computer hardware and storage resources. As a matter of fact, the detection efficiency of M-CenterNet has been slightly improved, although it brings a small decrease in accuracy. It achieves an excellent performance balance in speed and accuracy.

Secondly, to verify the influence of embedding the lightweight receptive field enhancement module (LRFEM) in different positions for the feature extraction ability, we design the ablation experiment of the HRM-CenterNet network without embedding LRFEM and with embedding LRFEM in different positions on performance. Table 3 shows the AP of plate fittings with large shape difference among yoke plate, adjusting board and hanging board, and the mAP of 8 types of fittings. The experimental results show that the module embedded in different positions all improves the ability of the model to capture different layers of context information. Performance improvement is the largest when the module is applied to 16 times downsampling feature map. When applied to the 32 times downsampling feature map, the receptive field of conventional convolution is large enough, so the accuracy is slightly improved after the module is embedded. In addition, as the number of layers embedded in the module deepens, the amount of calculation brought by the module increases gradually. Therefore, the module is applied to 16 times downsampling feature map in the HRM-CenterNet.

Thirdly, to verify the effectiveness of the lightweight receptive field enhancement module, we designed the performance experiment based on HRM-CenterNet using RFB module, RFB-s module, and our improved module, respectively. Table 4 shows that the amount of parameters is reduced by 40% compared with the RFB-s module after deleting part of the convolution layer. However, the detection accuracy is the same as before.

Table 3: The influence of LRFEM embedding position.

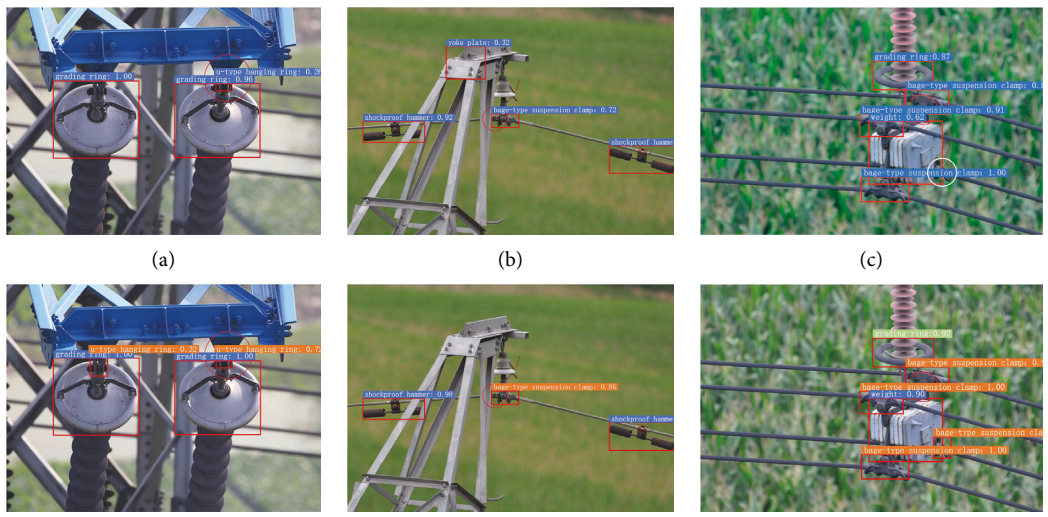
Position	AP/%			mAP
	Yoke plate	Adjusting board	Hanging board	
No LRFEM	69.4	77.2	57.2	79.0
With 4X	69.8	77.4	57.2	79.3
With 8X	71.3	79.5	58.9	79.7
With 16X	73.9	80.4	60.7	80.3
With 32X	73.5	80.0	58.6	79.8

Table 4: The influence of different RFB modules.

Methods	mAP/%	Parameter/M
RFB block	80.0	0.7
RFB-s block	80.3	1.0
Ours	80.3	0.6

Then, the results of ablation experiments for HRFF and LRFEM are shown in the last column of Table 1. When both methods are introduced, the detection accuracy of the fittings reaches the best. The detection accuracy is improved by 1.3% and 3.1% for LRFEM and HRFF, respectively, and it shows that the latter improves the accuracy of the model more significantly than the former.

Finally, we conduct comparative visual experiments. Figure 11 shows qualitative examples of HRM-CenterNet for fitting detection on the TFITS validation set. The above is the detection result of CenterNet, and the below is the result of HRM-CenterNet. False and missed detections have been marked with white boxes. For example, although the grading ring is detected in the CenterNet in Figure 11(a), the u-type hanging ring on the left grading ring was missed. In Figure 11(b), the tower material is mistakenly detected as a yoke plate. The bag-type suspension clamp blocked by the weight in the right corner in Figure 11(c) is not detected by the CenterNet. In Figures 11(d)–11(f), the detection results of HRM-CenterNet embedded with high-resolution feature fusion and multi-scale receptive field enhancement module have greatly improved these problems. Small fitting u-type hanging ring and bag-type suspension clamp due to dense occlusion are detected. In addition, the confidence of the bounding boxes has been significantly improved, representing the improvement of the feature extraction ability of



(a) (b) (c)

(d) (e) (f)

Figure 11: Qualitative examples of HRM-CenterNet. The above represents the detection results of CenterNet, and the below represents the detection results of HRM-CenterNet.

Table 5: The validation results on the PASCAL VOC dataset.

Methods	AP/%				mAP	FPS
	Potted plants	Cups	Chairs	Dining tables		
CenterNet-Res18	49.6	63.1	59.2	72.8	75.7	96.0
CenterNet-Res101	54.0	67.3	63.2	75.6	78.7	27.0
CenterNet-DLA	58.2	71.4	65.6	77.9	80.1	30.0
CenterNet-HG	58.7	71.5	67.2	78.2	81.5	9.5
M-CenterNet	53.1	66.4	62.0	76.6	78.2	35.5
HRM-CenterNet	59.3	71.6	67.1	79.1	81.6	26.0

the detectors. The experiments show that HRM-CenterNet reduces the probability of false and missed detection and verifies the effectiveness of the innovations.

4.5. Comparisons with Other Approaches. In order to compare with other state-of-the-art approaches, we trained our HRM-CenterNet in the TFITS Dataset and then submitted the results to the leaderboards. The results are shown in Table 1. The M-CenterNet and HRM-CenterNet have the smallest size, greatly reducing computer hardware and storage resources. For detection efficiency, although the HRM-CenterNet is lower than EfficientDet, it is 6.3% higher than EfficientDet in accuracy, and both of them have real-time performance. For detection accuracy, HRM-CenterNet ranks second, but its detection efficiency is more than twice that of rank 1 ExtremeNet. Although our HRM-CenterNet cannot achieve the best performance in every category, it achieves the best speed-accuracy trade-off among all the detectors shown in Table 1.

4.6. Experiments on PASCAL VOC 2007. We further compare our HRM-CenterNet with different backbone networks on PASCAL VOC 2007. The results are shown in Table 5. We selected four categories prone to missed and false detection, such as potted plants and chairs. The potted plants have large-scale changes and many small-scale targets, but the performance of HRM-CenterNet is much higher than that of M-CenterNet. The accuracy is improved by 6.3%, indicating that the LRFEM can better retain spatial semantic information, and the iterative aggregation feature fusion can make full use of this retained information. Compared with chairs, cups, dining tables, and other categories vulnerable to occlusion, HRM-CenterNet is superior to other algorithms. It shows that our network has a stronger feature extraction ability and good robustness. In summary, our HRM-CemterNet also achieves the best trade-off between accuracy and efficiency on PASCAL VOC 2007 dataset.

5. Conclusion

In conclusion, aiming to improve the performance of the anchor-free detectors for fitting object detection, we introduce feature enhancement and iterative aggregation to CenterNet. The detection accuracy is efficiently improved for introducing lightweight feature enhancement modules and high-resolution feature fusion, verified by the experimental results on the TFITS Dataset. The results on the TFITS and PASCAL VOC 2007 demonstrate that our HRM-CenterNet achieves the best speed-accuracy trade-off. Besides, HRM-CenterNet is suitable for deployment on the outdoor inspection platform. In further research, we will optimize the

detection performance of difficult samples. The next step of the research will be to extract richer features, smaller structures, and more efficient methods and then make further model performance improvements in embedded devices to achieve real-time fitting detection on the mobile terminal.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Science and Technology Project of State Grid Jibei Power Company Limited (no. B3018E200000).

References

[1] J. Toth and A. G. Jackson, "Smart view for a smart grid—unmanned Aerial Vehicles for transmission lines," in *Proceedings of the International conference on applied robotics for the power industry*, pp. 1–6, IEEE, Montreal, Canada, October 2010.

[2] W. Tong, J. Yuan, and B. Li, "Application of image processing in patrol inspection of overhead transmission line by helicopter," *Power System Technology*, vol. 34, no. 12, pp. 204–208, 2010.

[3] V. N. Nguyen, R. Jenssen, and D. Roverso, "Automatic autonomous vision-based power line inspection: a review of current status and the potential role of deep learning," *International Journal of Electrical Power & Energy Systems*, vol. 99, no. 2, pp. 107–120, 2018.

[4] Z. Zhao, H. Qi, Y. Qi, K. Zhang, Y. Zhai, and W. Zhao, "Detection method based on automatic visual shape clustering for pin-missing defect in transmission lines," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 9, pp. 6080–6091, 2020.

[5] A. Krizhevsky, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 54, no. 6, pp. 1097–1105, 2012.

[6] S. Fan, F. Zhu, S. Chen et al., "FII-CenterNet: an anchor-free detector with foreground attention for traffic object detection," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 1, pp. 121–132, 2021.

[7] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," 2019, https://arxiv.org/abs/1904.07850.

[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, IEEE, Columbia, Oh, USA, June 2014.

[9] H. Kuang, B. Wang, J. An, M. Zhang, and Z. Zhang, "Voxel-FPN: multi-scale voxel feature aggregation for 3D object detection from LIDAR point clouds," *Sensors*, vol. 20, no. 3, pp. 704–710, 2020.

[10] K. Zhang, K. Zhao, and X. H. Feng, "HRM-CenterNet: a high-resolution real-time fittings detection method," in *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 564–569, IEEE, Melbourne, Australia, October 2021.

[11] D. Hang, X. Qiu, and C. Cao, "Algorithm of vibration damper detection combined with aggregation channel and complex frequency domain features," *Computer Technology and Development*, vol. 36, no. 3, pp. 147–151, 2020.

[12] S. Han, R. Hao, and J. Lee, "Inspection of insulators on high-voltage power transmission lines," *IEEE Transactions on Power Delivery*, vol. 24, no. 4, pp. 2319–2327, 2009.

[13] Q. Zhou, X. Wu, S. Zhang, B. Kang, Z. Ge, and L. Jan Latecki, "Contextual ensemble network for semantic segmentation," *Pattern Recognition*, vol. 122, pp. 108290–108301, 2022.

[14] Q. Zhou, Y. Wang, Y. Fan et al., "AGLNet: towards real-time semantic segmentation of Self-driving images via attention-guided lightweight network," *Applied Soft Computing*, vol. 96, no. 11, pp. 106682–106694, 2020.

[15] Q. Zhou, W. Yang, G. Gao et al., "Multi-scale deep context convolutional neural networks for semantic segmentation," *World Wide Web*, vol. 22, no. 2, pp. 555–570, 2019.

[16] Q. Zhou, J. Wang, and J. Liu, "RSANet: towards real-time object detection with residual semantic-guided attention feature pyramid network," *Mobile Networks and Applications*, vol. 26, no. 4, pp. 77–87, 2021.

[17] Z. Zheng, P. Wang, and W. Liu, "Distance-IoU loss: faster and better learning for bounding box regression," in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 12993–13000, New York, NY, USA, February 2020.

[18] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, IEEE, Las Vegas, NV, USA, June 2016.

[19] W. Liu, D. Anguelov, and D. Erhan, "SSD: single shot multibox detector," in *Proceedings of the European conference on computer vision*, pp. 21–37, Springer, Amsterdam, Netherlands, October 2016.

[20] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7263–7271, IEEE, Hawaii, Hl, USA, July 2017.

[21] J. Redmon and A. Farhadi, "Yolov3: an incremental improvement," 2018, https://arxiv.org/abs/1804.02767.

[22] B. Benjdira, T. Khursheed, and A. Koubaa, "Car detection using unmanned aerial vehicles: comparison between Faster R-CNN and Yolov3," in *Proceedings of the 2019 1st International Conference on Unmanned Vehicle Systems-Oman (UVS)*, pp. 1–6, Muscat, Oman, February 2019.

[23] A. Bochkovskiy, C. Wang, and H. Liao, "Yolov4: optimal speed and accuracy of object detection," 2020, https://arxiv.org/abs/2004.10934.

[24] S. Liu, L. Qi, and H. Qin, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8759–8768, IEEE, Salt Lake City, UT USA, June 2018.

[25] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R.-C. N. N.: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[26] C. Fu, W. Liu, and A. Ranga, "DSSD: deconvolutional single shot detector," 2017, https://arxiv.org/abs/1701.06659.

[27] R. Girshick, "Fast R-CNN," in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, IEEE, Santiago, Chile, May 2015.

[28] K. He and G. Gkioxari, "Mask R-CNN," in *Proceedings of the IEEE international conference on computer vision*, pp. 386–397, IEEE, Venice, Italy, October 2020.

[29] Z. Cai and N. Vasconcelos, "Cascade R-CNN: delving into high-quality object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6154–6162, Salt Lake City, UT, USA, June 2018.

[30] C. Zhu, Y. He, and M. Savvides, "Feature selective anchor-free module for single-shot object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 840–849, Long Beach California, USA, June 2019.

[31] H. Law and J. Deng, "CornerNet: detecting objects as paired keypoints," in *Proceedings of the European conference on computer vision*, pp. 734–750, Munich, Germany, September 2018.

[32] X. Zhou, J. Zhuo, and P. Krahenbuhl, "Bottom-up object detection by grouping extreme and center points," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 850–859, California, USA, June 2019.

[33] K. Duan and S. Bai, "CenterNet: keypoint triplets for object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 6569–6578, Seoul, Korea, October 2019.

[34] I. Ahmed, M. Ahmad, J. J. P. C. Rodrigues, and G. Jeon, "Edge computing-based person detection system for top view surveillance: using CenterNet with transfer learning," *Applied Soft Computing*, vol. 107, no. 6, pp. 107–114, 2021.

[35] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: fully convolutional one-stage object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 9627–9636, Seoul, Korea, October 2019.

[36] Y. Zhai, D. Wang, M. Zhang, J. Wang, and F. Guo, "Fault detection of insulator based on saliency and adaptive morphology," *Multimedia Tools and Applications*, vol. 76, no. 9, pp. 12051–12064, 2017.

[37] Y. Zhai, H. Cheng, R. Chen, Q. Yang, and X. Li, "Multi-saliency aggregation-based approach for insulator flashover fault detection using aerial images," *Energies*, vol. 11, no. 2, pp. 340–349, 2018.

[38] Y. Hao, J. Wei, X. Jiang et al., "Icing condition assessment of in-service glass insulators based on graphical shed spacing and graphical shed overhang," *Energies*, vol. 11, no. 2, pp. 318–328, 2018.

[39] Y. Liu, J. Li, and W. Xu, "A method on recognizing transmission line structure based on multi-level perception," in *Proceedings of the International Conference on Image and Graphics*, pp. 512–522, Shanghai, China, September 2017.

[40] Y. Tang, J. Han, and W. Wei, "Research on part recognition and defect detection of transmission line in deep learning," *Electronic Measurement Technology*, vol. 41, no. 6, pp. 60–65, 2018.

[41] Q. Gao and Q. Lian, "Research on target detection of the insulator in the aerial image," *Electrical Measurement & Instrumentation*, vol. 21, no. 6, pp. 119–123, 2019.

[42] Z. Zhao, A. Jiang, and Y. Qi, "Fittings detection of transmission line image based on SSD model embedded with occlusion relation module," *Journal of Intelligent Systems*, vol. 14, no. 8, pp. 343–348, 2020.

[43] Y. Qi, A. Jiang, and Z. Zhao, "Fittings detection method in patrol images of transmission line based on improved SSD," *Electrical Measurement & Instrumentation*, vol. 15, no. 4, pp. 656–662, 2020.

[44] Z. Zhao and Y. Li, "Typical fittings detection method with faster R-CNN combining KL divergence and shape constraints," *High Voltage Engineering*, vol. 46, no. 9, pp. 123–125, 2020.

[45] R. Zhao and G. Zhao, "A real-time fault detection method for high voltage transmission line based on CenterNet Improved Algorithm," *Computer Engineering and Applications*, vol. 34, no. 8, pp. 34–37, 2020.

[46] M. Zhao and Y. Lu, "Object detection technology in the complex environment based on CenterNet algorithm," *Journal of China Academy of Electronics*, vol. 64, no. 6, pp. 654–660, 2021.

[47] A. Howard, M. Sandler, and G. Chu, "Searching for mobilenetv3," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1314–1324, Seoul, Korea, October 2019.

[48] F. Chollet, "Xception: deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1251–1258, Hawaii, Hl, USA, July 2017.

[49] S. Liu and D. Huang, "Receptive field block net for accurate and fast object detection," in *Proceedings of the European Conference on Computer Vision*, pp. 385–400, Munich, Germany, September 2018.

[50] K. Sun, B. Xiao, and D. Liu, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5693–5703, California, USA, June 2019.

[51] T. Lin, P. Goyal, and R. Girshick, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988, Venice, Italy, October 2017.

[52] M. Tan, R. Pang, and Q. Le, "Efficientdet: scalable and efficient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10781–10790, Seattle, WA USA, June 2020.