IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 41, No. 2, April, 2024, pp. 919-927
Journal homepage: http://iieta.org/journals/ts

Dual-Scale Dataset-Based Intelligent Recognition of Power Equipment for Enhanced Digital Grid Planning

Xichun Feng[1*], Jinglin Han[2], Yang Liu[1], Ruosong Hou[1], Tieliang Li[3]

[1] State Grid Hebei Economic Research Institute, Hebei 050000, China
[2] State Grid Hebei Electric Power Company, Hebei 050000, China
[3] State Grid Hengshui Electric Power Company, Hebei 053000, China

Corresponding Author Email: fengxichun144@gmail.com

https://doi.org/10.18280/ts.410232

Received: 16 December 2023
Revised: 15 March 2024
Accepted: 31 March 2024
Available online: 30 April 2024

Keywords:
power equipment identification, deep learning, object detection, dual-scale datasets

ABSTRACT

In the evolution of digital power grids, technologies such as artificial intelligence and digital twins have emerged as foundational elements. A critical step in this evolution involves creating digital avatars of physical power equipment, enabling precise classification crucial for digital twin integration. Traditional algorithms for equipment recognition face significant challenges due to discrepancies in equipment sizes, and environmental factors like bright light and haze, which substantially degrade detection performance. This study introduces a novel single-stage recognition method that employs staged training on a dual-scale dataset, specifically designed to address these challenges. Equipment images are categorized into large-scale and small-scale sets to mitigate issues arising from size disparities. Furthermore, a comprehensive dataset featuring multiple scales, angles, and lighting conditions is compiled, enhancing the model's generalizability and robustness. The proposed method incorporates a feature extraction module, a feature fusion network, and environment context modeling, which are trained separately on the large-scale and small-scale datasets. During testing, outputs from the dual-scale models are integrated. Comparative experiments demonstrate that the proposed method requires only one-third the parameters of the SSD algorithm, yet operates at a detection speed of 52.15 frames per second (fps). Despite its lightweight structure, the algorithm achieves an impressive mean Average Precision (mAP) of 92.13%, effectively reducing false and missed detections. This performance signifies a marked improvement in stability and robustness over existing single-stage detection methods, particularly in complex natural environments.

1. INTRODUCTION

As industries and agriculture undergo automation, the demand for electricity has surged, exacerbating the challenge of efficiently allocating power resources. To ensure the stable operation of electrical equipment, effective power grid planning is paramount. Modern power grid planning aims to construct a resilient, intelligent power grid. Traditional planning methods rely heavily on manual labor, while digital planning, characterized by information-based, automated, and intelligent approaches, offers a promising alternative [1]. In recent years, digital technology has permeated various sectors of the economy and society. Notably, artificial intelligence and digital twins have emerged as crucial components for establishing a digital power grid [2]. A digital power grid necessitates the digitization of the entire process, encompassing all elements and business aspects of the power grid, to create a seamless virtual digital world mirroring the physical one. Leveraging digital information, novel business models and optimization methods can be devised to enhance the operation and management of the power grid [3].

The cornerstone technologies of a digital twin system include modeling, simulation, and data fusion-based digital threads. Establishing a corresponding "digital replica" of real power equipment along the physical power lines is a prerequisite for a digital twin power grid. Moreover, effective management and classification of this "digital replica" are essential for conducting precise simulations and optimizations. Intelligent recognition algorithms play a pivotal role in automatically tagging and classifying collected device image information. However, developing robust power equipment recognition algorithms poses numerous challenges. These algorithms must accurately detect various objects, ranging from large devices like wires, utility poles, and transformers to smaller components like insulators and crossbars. Disparities in scale among different power equipment components further complicate accurate detection. Additionally, complex environmental factors such as strong light and haze can distort object features, necessitating solutions to maintain high detection accuracy under varying conditions. Furthermore, for deployment on edge data acquisition devices such as robots and drones, detection algorithms must prioritize lightweight designs and meet real-time performance requirements.

Presently, research on power equipment target detection and recognition primarily revolves around traditional detection

algorithms and deep learning-based approaches. Traditional algorithms rely on manually designed features for template matching, followed by feature classification and regression to achieve object recognition [4]. While these algorithms offer fast detection speeds, they often struggle with robustness in practical applications due to complex environmental factors, leading to issues like false positives and false negatives. In contrast, deep learning algorithms leverage hierarchical feature extraction and learning capabilities to achieve breakthroughs in object detection performance. These methods automatically learn advanced features from massive datasets, significantly improving recognition accuracy. Many studies have applied deep convolutional neural networks (CNNs) to power equipment detection, reflecting the general trend towards using intelligent detection methods to maintain power grid systems. For instance, Jiang et al. [5] propose a recognition model for substation equipment images based on Mask-RCNN [6] and Faster-RCNN [7], highlighting the need for enhancements to achieve better results in specific application scenarios. Similarly, Lin et al. [8] introduce a transmission line inspection image detection algorithm based on an improved Faster-RCNN, addressing challenges related to recognition accuracy and real-time detection. Despite these advancements, certain limitations persist, such as slow recognition speeds, large model sizes, and challenges in optimizing detection accuracy under various environmental conditions. Wan et al. [9] use the position information of largerscale power equipment to correct the probability of small components based on the position-related relationship between small and large components of transformers, and improve the recognition accuracy of the small objects, but a large number of conditional judgments need to be artificially set in the postprocessing process, and the generalization ability is weak. Chen et al. [10] improved the single-stage object detection algorithm, RFBNet [11], use a lightweight backbone network to improve the model's detection speed, but the detection accuracy of small-scale equipment still needs to be further improved. Xiong et al. [12] add a de-fogging algorithm to the power equipment detection algorithm for data preprocessing to improve the recognition accuracy of the detection algorithm in foggy weather, but the detection accuracy in other complex environments such as strong light and night scenes still needs further optimization. Due to the slow speed and large model parameter size of two-stage object detection algorithms such as RCNN [13] and Fast RCNN [14], they cannot meet the real-time detection requirements.

In this paper, we present an improved single-stage multiscale object detection algorithm designed to address these challenges. Leveraging lightweight backbone networks, context mining, and feature fusion technology, our algorithm enhances detection accuracy and speed while significantly reducing model parameter size. We constructed a large-scale recognition dataset encompassing multiple categories of power equipment for training purposes. To tackle the issue of poor recognition caused by extreme differences in target scales, we divided device images into two scales and employed multi-stage training on data from different scales. Various data augmentation techniques were applied to enhance the algorithm's generalization ability and robustness against diverse weather and lighting conditions. During testing, the model was trained on both large-scale and small-scale device datasets, and the detection results were combined to establish an end-to-end multi-scale power equipment recognition pipeline. Our proposed method achieved a mAP of 92.13% and a recognition speed of 52.15 fps on the test set. Extensive experiments demonstrate the efficacy of our multiscale recognition method in power equipment detection, with potential applications in constructing digital twin power grid equipment tags.

2. RELATED WORKS

Deep learning models have recently been broadly applied throughout the whole area of computer vision, including generic and task-specific object detection. Detectors typically use CNNs to extract features from input images, classify objects, and locate them. A few important contributions have been made since the first CNN-based object detector, R-CNN [13], was proposed. In summary, the object detectors can be divided into anchor-based and anchor-free methods.

2.1 Anchor-based detectors

Anchor-based detectors inherit ideas from the sliding window and region proposals. In the anchor-based pipeline, anchors serve as the fundamental element for classification and location refinement. They regard objects as a partial area of a picture and use anchors to determine the boundary and scope of the area. During training, the network gradually refines the scale and position of these anchors. Based on anchors, a model has an initial optimization point, which promotes the neural network to converge. Today's anchor-based methods can be divided into two-stage and one-stage detectors. The former defines detection as the process of "going from coarse to fine," while the latter defines detection as "one-step completion." R-CNN, as the first successful anchor-based object detector, enumerates a large set of candidates as region proposals in the first stage and classifies the cropped candidate boxes using a deep CNN. Then, SPP-Net [15], Fast R-CNN, Faster R-CNN, and other approaches based on R-CNN are proposed, which have improved efficiency and accuracy to a certain extent. Two-stage methods rely on region proposals for classification and regression, while one-stage detectors complete the classification and regression with one network. SSD [16] sets a series of prior anchors before training and uses multi-scale feature maps, which significantly improve single-stage detector accuracy and become the baseline for subsequent approaches. Recently, YOLOv3 [17], RetinaNet [18], EfficientDet [19], and other detectors have been proposed. These are all anchor-based, one-stage detectors with good performance.

2.2 Anchor-free detectors

Anchor-free detectors do not use prior anchors or region proposal networks (RPN) and solve object detection as a keypoint estimation task. The emergence of CenterNet [20], CornerNet [21], and other anchor-free detection methods has brought new inspiration to object detection. The rise of full convolution networks [22] (FCN) provides a new paradigm that solves the object detection problem through the local response on the feature map. The following works have shown that the performance of anchor-free detectors can be equal to that of anchor-based detectors. CornerNet detects objects as a pair of keypoints, the top-left corner and bottom-right corner of the bounding box. At the detection head, a convolutional layer is used to generate a heatmap and predict an embedding vector for each detected corner. CenterNet represents objects

with a single point at the bounding box center. Other properties, including size, dimension, and orientation, are regressed from predictions. Built on down-scale feature maps, FCOS [23] constructs a set of detection heads sharing parameters to deal with objects of various scales. Each head detects the center of a bounding box, and an offset branch is built to predict the deviation of the estimated center. YOLOX [24], which is an improved version of YOLOv3, modifies the baseline to be anchor-free. Two parallel prediction branches are integrated into the detection head for classification and regression. Such improvements achieve better performance compared to YOLOv3.

3. METHOD

3.1 Overview

Existing single-stage object detection algorithms (e.g., YOLO [25] and SSD) pre-set prior boxes with different aspect ratios on the feature maps. A prior box will be classified as an object by the network if the overlap ratio between the prior box and the ground truth box reaches 0.5. The network gradually fine-tunes the sizes of the prior boxes through training to fit the ground truth object. However, due to the scale of the small objects, which have a relatively small range on the feature map, the corresponding prior boxes will have a larger deviation from the ground truth boxes. Thus, it is difficult for the network to filter the small objects and take them into the training process, which will also lead to poor detection performance for small objects during testing. The scales of different power equipment vary greatly. Large-scale equipment includes transformers, utility poles, etc.; densely packed small-scale equipment includes crossbars and insulators. As shown in Table 1, the scale of large equipment is tens or hundreds of times that of small equipment. The image of equipment with different scales is shown in Figure 1. It can be seen that the overall contour of the large-scale equipment is clear. But it is difficult to see the details of the small-scale equipment's details are difficult to see. Simply increasing the resolution of small-scale equipment is not a suitable solution. Because the large-scale equipment in the image will only have local information, The input image resolution of the SSD algorithm is 300×300 or 512×512. Small-scale equipment will lose more details at this resolution, making detection difficult.

Table 1. The ratio of object area to image size

Object	Utility Pole	Transformer	Crossbar	Insulator
Area Ratio	1.483%	0.663%	0.068%	0.017%

Figure 1. Scale comparison between large equipment and small equipment: the green box is the area of large equipment, and the red box is the area of small equipment

To solve the problem of low detection accuracy caused by differences in the scales of electric equipment, we divide the dataset into dual-scale datasets (large-scale and small-scale) based on the size of the equipment. In this paper, we propose a detection approach based on training on dual-scale datasets and an improved single-stage multi-scale object detection algorithm. The network is trained on both the large-scale and small-scale datasets separately. The outputs of the two models are merged during testing, followed by non-maximum suppression (NMS), to generate unified detection results. Figure 2 shows the complete pipeline of the proposed framework.

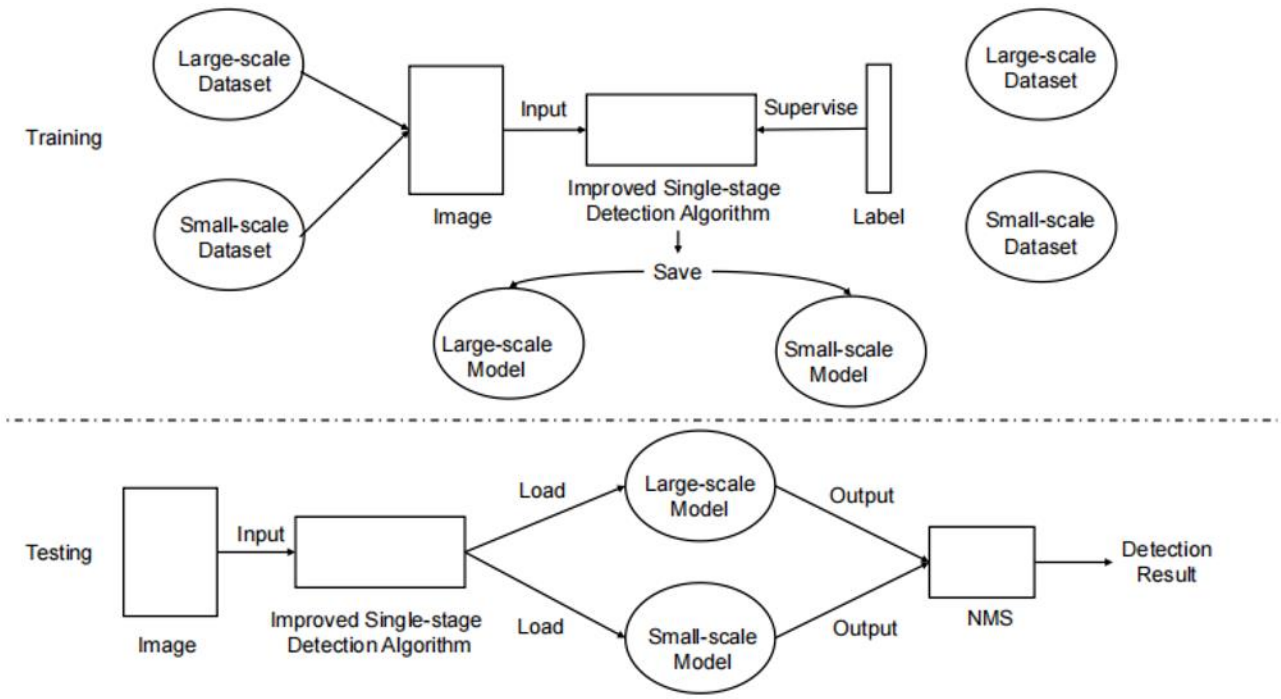

Figure 2. Overall pipeline

The detection algorithm is single-stage with a feature extraction module, a feature fusion network, and a detection head. The feature fusion network combines multi-scale features with global context for object detection. To facilitate model deployment on edge devices such as drones, our backbone network replaces VGGNet [26] with the more efficient and lightweight EfficientNet [27]. We improve the original feature pyramid structure with multiple levels of feature fusion to balance the network's ability to extract semantic features and locate objects. Consequently, our network is adaptable to different sizes of equipment data in real-world scenarios. The detection neck also incorporates an attention-like global semantic information module, which enhances the network's focus on the target by exploiting spatial domain context information and weakens interference from cluttered backgrounds.

3.2 Dual-scale dataset

We classify data based on the equipment's true scale. Large-scale data refers to large-scale power equipment such as transformers and utility poles; small-scale data consists of dense small equipment such as crossbars and insulators. To balance the algorithm's recognition ability for objects of different scales, we build two datasets of large and small-scale power equipment, respectively, and train the network separately. This strategy solves the problem of insufficient training of small objects caused by mismatches between prior boxes and ground truth, thus improving the accuracy of small-scale device detection.

We use real images of power equipment on transmission lines as the data source, which covers multiple scales, angles, and lighting conditions. The dataset includes four types of equipment: utility poles, transformers, insulators, and crossbars. The resolution of the actual shooting image is 3000×4000, and the equipment scale ratio is determined based on the area ratio of the equipment in the image. We classify poles and transformers as large-scale equipment, and insulators and crossbars as small-scale equipment. The dataset is annotated using the LabelImg tool, and the annotation format follows the PASCAL VOC 2007 dataset format. To facilitate network training, the resolution of the images in the training set is fixed at 300×300. Since large-scale equipment occupies the entire image frame of the actual image, the actual shooting image is scaled to 225×300 and padded with zeros around the edges to obtain the large-scale training set. Actual shooting images containing small-scale equipment are selected, and a series of sub-images are generated by traversing the image using a sliding window with a resolution of 300×300 and a stride of 15×15. The sub-images are cropped and labeled to create a small-scale dataset. Table 2 shows the number of objects for each class in the dataset. The dataset is randomly sampled and divided into training and testing sets, with a ratio of 9:1 for both large and small-scale datasets.

Table 2. Statistics of various objects in the dataset

Object	Utility Pole	Transformer	Crossbar	Insulator
Number	2341	549	1836	7044

In actual outdoor scenarios, there are complex natural environmental influences, including noise, backlighting, changing weather conditions, as well as complex scale and angle changes. It is difficult to gain rich and realistic data with manual photography, which limits the detection accuracy of algorithms in extreme situations. To improve the generalization ability of the model, enhance the robustness of the recognition algorithm, and strengthen the performance under actual scenarios, we follow the data optimization method explored by Zoph et al. [28] and have applied various data augmentation techniques to both datasets.

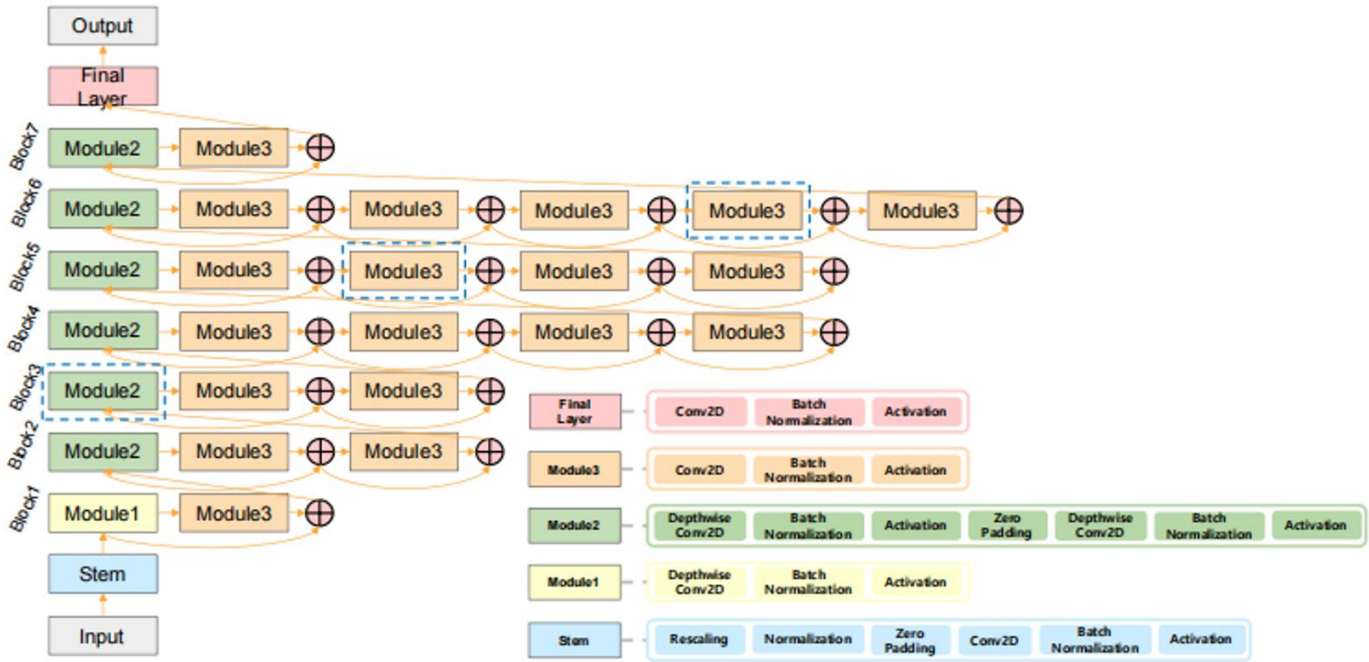

Figure 3. The network structure of EfficientNet-b3

3.3 Single-stage object detection algorithm

To optimize detection speed, edge devices require lightweight and efficient detection algorithms. The SSD algorithm has demonstrated commendable detection performance on public datasets, owing to its single-stage

structure, which facilitates seamless integration across diverse platforms. In this paper, we introduce an enhanced object detection approach that integrates multiscale features and global context into the SSD algorithm. The network architecture of our algorithm is depicted in Figure 3. We improved the backbone network with EfficientNet-b3. The VGG network is based on the traditional stacked convolutional layer structure. But the network depth of VGGNet is not enough, and it uses more parameters. The EfficientNet is constructed by neural architecture search (NAS), achieving better accuracy and efficiency while expanding the network scale. By searching for optimization, the depth and width of each convolutional layer are determined, and then the modules are stacked to maximize the feature extraction ability of the model with limited model parameters. Introducing EfficientNet-b3 in the feature extraction part can significantly reduce the number of parameters and improve the feature extraction ability, enabling the model to increase feature extraction performance and be greatly lightweight, so that the model can be deployed on edge devices with limited device storage space. EfficientNet b3 is composed of seven modules in series. Each module contains a series of convolutional, batch normalization, and activation layers. The network structure diagram of EfficientNet-b3 is shown in Figure 3. We select the output feature maps from the 7th, 17th, and 25th convolutional modules for object recognition and detection, and the scale of these feature maps is 38×38, 19×19, and 10×10, respectively. For predicting the position and size of detection boxes, we select shallow feature maps, which achieve good results.

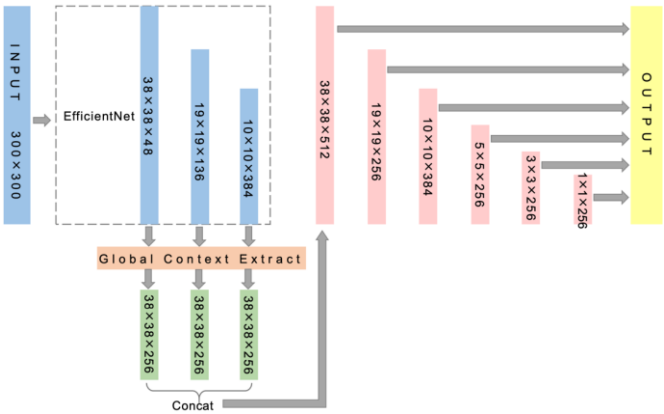

Figure 4. The network structure of the proposed single-stage algorithm

Object detection is a multifaceted task, encompassing two crucial components: classification and localization. The classification aspect involves learning high-level semantic information to establish invariant feature representations. Meanwhile, localization aims to pinpoint changes in object location and scale, necessitating an understanding of the interplay between the object and its background, as well as the relationships between local and global features. Achieving this necessitates the acquisition of rich equivariant features within the lower-level feature maps. Consequently, an effective detector must simultaneously learn both high-level semantic features and shallow, locally equivariant features. CNNs naturally organize themselves into a hierarchical feature pyramid. However, in the case of Single Shot MultiBox Detector (SSD), it independently detects objects of varying scales on different branches of these feature maps, overlooking the correlations between feature maps at various levels. This leads to an imbalance that affects the detection performance, causing a disparity between semantic and detailed information across different levels of feature maps. To address this issue and ensure more efficient utilization of extracted features, it becomes essential to enhance the semantic content of the shallow feature maps. In this research paper, we propose a novel approach: concatenating the feature maps from different levels for feature fusion. This concatenation approach offers several advantages. Notably, it obviates the need for converting feature maps from different levels to the same channel, a requirement in some other fusion methods. Consequently, it enhances the flexibility of fusing feature maps while reducing computational resource consumption compared to element-wise summation.

In our approach, we incorporate three feature maps produced by the backbone network and apply upsampling to standardize the feature map size to 38×38 while preserving a channel size of 256. Subsequently, we interconnect these feature maps to create a unified 38×38×768 feature map, as illustrated in Figure 4. This novel methodology ensures the seamless integration of both high-level semantic information and shallow, locally equivariant features, resulting in enhanced object detection performance.

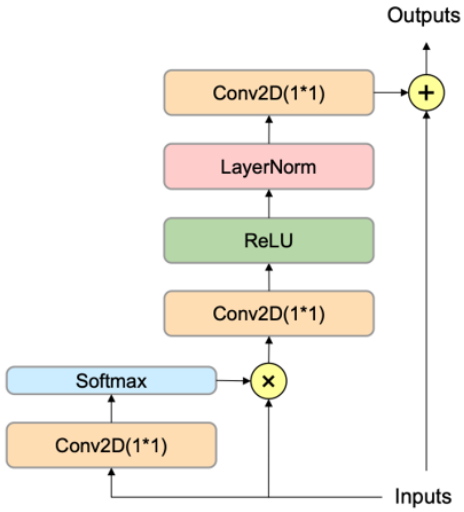

Figure 5. The network structure of GCBlock

The convolution operation utilizes local information to calculate the target pixel, and the receptive field is determined by the size of the convolution kernel. Due to the inability to introduce global information, traditional convolution has limitations in prediction. Currently, there are some straightforward methods to alleviate this problem, such as using larger convolution filters or building deeper networks. However, these methods do not significantly improve the results. The receptive fields of specific layers are still limited even with feature fusion, which increases the computational costs. The SSD algorithm does not consider the impact of environmental context on local objects, greatly limiting its ability to recognize small-scale, occluded, and low-resolution objects. There is a certain positional correlation between different pieces of power equipment. For example, in large-scale datasets, transformer equipment is usually installed between two utility poles; in small-scale datasets, there are adjacent positions between insulators and crossbars. Introducing context information into the model can implicitly model the positional correlation between different pieces of

equipment and avoid missed or false detections. However, not all environmental context information is beneficial to improving object detection performance; adding meaningless background noise may even impair detection performance. Therefore, identifying useful context information is necessary. We introduce a global context block (GCBlock) for modeling context semantics at specific query positions. It can help the model assign different weights to remote positions of the input, extract critical information, and make more accurate judgments without causing significant computational and storage costs. As shallow feature maps contain more context information, the first three feature maps output by the backbone network are input into the GCBlock for context information extraction to obtain more accurate positional correlation modeling. Figure 5 shows the network structure of the global context block.

During training, if the intersection over union (IoU) between the prior box and the ground truth box is greater than 50%, it can be considered that the prior box matches the true target. The network simultaneously performs object classification and box regression on the matched prior boxes, and the total loss function is calculated as Eq. (1).

$$L(x,c,l,g) = \frac{1}{N}\Big(L_{cls}(x,c) + \alpha L_{reg}(x,l,g)\Big), \tag{1}$$

where, $L_{cls}(x,c)$ is the classification loss, $L_{reg}(x,l,g)$ is the regression loss, N is the number of matched detection boxes, and α is a weighting factor, which is usually set to 1. We use the cross-entropy loss for classification, as in Eq. (2).

$$L_{cls}(x,c) = -\sum_{i\in Pos}^{N} x_{ij}^{k} \log\Big(\emptyset\big(c_i^k\big)\Big) - \sum_{i\in Neg}^{N} \log\Big(\emptyset\big(c_i^0\big)\Big), \tag{2}$$

where, $x,k,i,j \in \{0, 1\}$ is the true class label of the object, using one-hot encoding, and k is the number of object classes. $\emptyset$ is the Softmax function, which gives the probability that the target belongs to a certain class. $L_{reg}(x,l,g)$ regresses the detection frame position and size using smooth L1 loss to calculate the relative distance between the detection frame and the real frame calculated as Eq. (3).

$$lL_{reg}(x,l,g) = \sum_{i\in Pos}^{N}\sum_{m\in Box} x_{ij}^{k}\varphi\big(l_i^m - \overline{g_j^m}\big), \tag{3}$$

$$Box = \{cx, cy, w, h\}, \tag{4}$$

$$\overline{g_j^{cx}} = \frac{g_j^{cx} - d_i^{cx}}{d_i^{w}}, \tag{5}$$

$$\overline{g_j^{cx}} = \frac{g_j^{cy} - d_i^{cy}}{d_i^{h}}, \tag{6}$$

$$\overline{g_j^{w}} = \log\left(\frac{g_j^{w}}{d_j^{w}}\right), \tag{7}$$

$$g_j^{h} = log\left(\frac{g_j^{h}}{d_j^{h}}\right), \tag{8}$$

where, l_i^m is the relative offset of the m-parameters of the i-th detection frame, $\overline{g_j^m}$ is the offset between the j-th real labeled box and the prior box, g_m^j and d_i^m correspond to the m-parameters of the real labeled box and the prior box, respectively.

4. EXPERIMENT AND RESULT ANALYSIS

4.1 Experimental platform and parameters

The experimental environment for training and testing is Ubuntu 20.04, NVIDIA GeForce RTX 2080 Ti, and Intel i9-9900K. The transfer learning model is pre-trained on ImageNet. We first train the large-scale dataset with 120k iterations. Then we stop and train the small-scale dataset with 160k iterations. Each image in the training set has a size of 300×300, and the batch size is set to 16. We use Adam as the optimizer, with an initial learning rate of 0.001 and a parameter of 0.9. We also use the poly decay strategy to dynamically adjust the learning rate, as shown in Eq. (9).

$$lr = lr_0 \times \left(1 - \frac{epoch}{N}\right)^{\gamma}, \tag{9}$$

where, lr represents the current learning rate, lr_0 represents the initial learning rate, epoch represents the current iteration round, N represents the maximum number of iteration rounds, and γ is a hyperparameter that is set to 0.9.

4.2 Experimental results and evaluation

4.2.1 Detection results

We train and store the model using both large-scale and small-scale datasets, employing phased detection during testing. Real-world images serve as inputs to the network. Initially, these images undergo large-scale object detection, followed by feeding the detection results into the small-scale detection network to produce final detection boxes. Subsequently, NMS is applied. We evaluate the recognition performance using power equipment images under diverse conditions, encompassing environmental factors like occlusion, shadows, and backlighting, as well as image acquisition variables such as shooting angle, scale, and rotation.

Figure 6. Detection results of real image under different weather, light and natural environment conditions, and different angles and scales

The detection results are shown in Figure 6. All large-scale objects are detected completely, without false or missed detections. The small-scale target detection results are pretty good, though some objects are not detected due to their incomplete appearance and insufficient features caused by occlusion. Since the small-scale dataset retains the appearance details of small devices well and the small and large-scale models are trained independently, the detector does not suffer from cross-scale interference. Therefore, our network allows the small-scale dataset to be sufficiently learned, and effectively solves the problem of poor small-scale target detection performance in single-stage detection algorithms.

4.2.2 Comparative experiments

In the field of object detection, accuracy and recall are generally used as quantitative indicators to evaluate recognition accuracy, and fps is used to evaluate the model's inference speed. For recognition accuracy, precision is defined as the proportion of correctly identified objects in all positive samples, which measures whether the algorithm can distinguish between positive and negative examples. Recall is defined as the proportion of true objects that can be correctly detected by the model in all true objects, which measures whether the algorithm can find all positive examples. In an actual detection task, it is difficult to balance precision and recall. So, we use the AP value, which combines the calculated precision and recall, to evaluate the algorithm's comprehensive performance, as shown in Eq. (12):

$$Precision = \frac{TP}{TP + FP} \times 100\%, \tag{10}$$

$$Recall = \frac{TP}{TP + FP} \times 100\%, \tag{11}$$

$$AP = \int_0^1 PdR, \tag{12}$$

where, *TP* is the number of correctly detected objects by the model, *MP* represents the number of incorrectly detected objects by the model, and *FN* is the number of correctly identified objects that the model missed. To evaluate the model's recognition performance for each target class in multiple classes *N*, the *mAP* is used, as shown in Eq. (12):

$$mAP = \frac{1}{N}\sum_{k=1}^{N} AP_k \tag{13}$$

We perform a series of comprehensive experiments to assess the algorithm's performance, comparing it against other single-stage object detection algorithms, namely SSD [16], YOLOv3 [17], and RetinaNet [18], as depicted in Figures 7-9. Our findings reveal that the algorithm proposed in this paper exhibits superior average detection accuracy compared to its counterparts. Additionally, it demonstrates faster inference speed and boasts a smaller model parameter size, rendering it particularly well-suited for deployment on data collection devices with constrained storage capacity.

To address these limitations and further enhance algorithm performance, several potential avenues for improvement can be explored. These may include refining the algorithm's feature extraction capabilities to better capture intricate details across different scales and orientations. Additionally, augmenting the training dataset with a diverse range of images that closely mirror real-world scenarios can aid in improving the algorithm's generalization ability. Furthermore, fine-tuning model parameters and exploring novel optimization techniques could contribute to further enhancing both detection accuracy and inference speed.

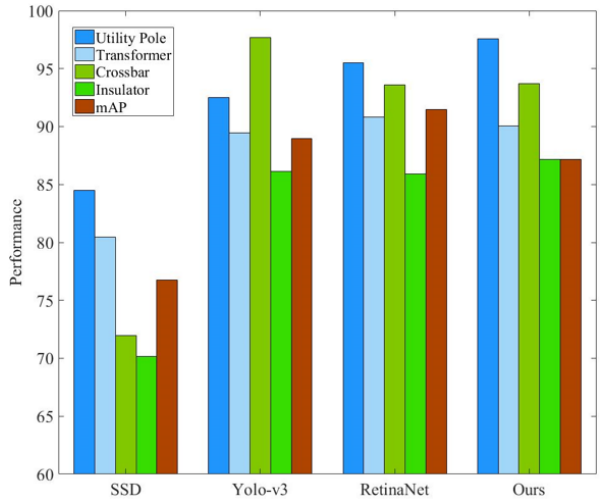

Figure 7. The detection accuracy of different algorithm

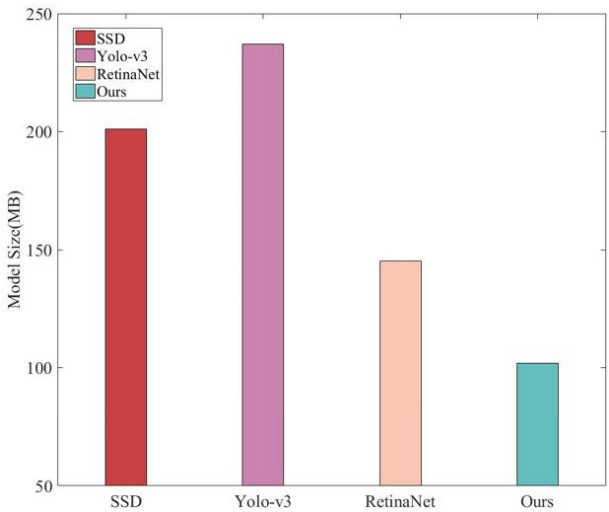

Figure 8. The model size of different model

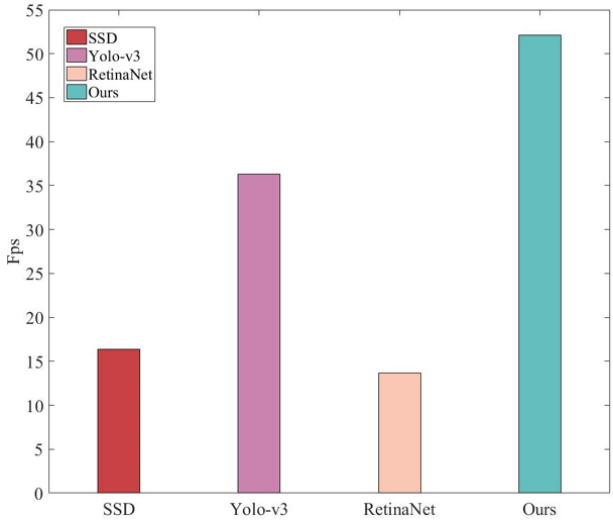

Figure 9. The fps of different model

Overall, our experiments provide valuable insights into the strengths and limitations of the proposed algorithm, paving the way for future research directions aimed at advancing single-stage object detection methodologies.

4.2.3 Ablation experiment

The proposed method offers significant enhancements over the SSD algorithm in areas such as the feature extraction module and feature fusion module, as confirmed by our ablation studies. Notably, substituting VGGNet with EfficientNet-b3 in our model markedly bolsters its feature extraction capabilities, which in turn substantially boosts inference performance. This is attributable to EfficientNet-b3's ability to optimize the parameter size of each module through a search strategy, enhancing performance without significantly increasing the parameter count. Specifically, adopting EfficientNet-b3 as the backbone network results in a remarkable 321% acceleration in inference speed compared to using VGGNet. Moreover, the integration of our feature fusion module allows for a more effective utilization of bottom-level feature maps' equivariant information, significantly improving the model's precision in object localization. Our feature fusion approach not only outperforms other methods in terms of efficiency but also achieves this with minimal impact on inference speed. Before feature fusion, our algorithm models the environmental context, subtly increasing parameter count but significantly boosting recognition accuracy for dense and occluded objects. By leveraging the spatial relationship between targets and the overall environment, our method prioritizes critical global features, offering a distinct advantage in identifying stationary objects. Omitting the feature pyramid and global context modules notably hinders network convergence speed, underscoring our algorithm's superior adaptability and stronger fit compared to the SSD algorithm. This makes it highly applicable across diverse scenarios (Table 3).

Table 3. Ablation experiment

VGGNet	EfficientNet-b3	Feature Fusion	GCBlock	Large-Scale	Small-Scale	Overall	fps
√				80.28%	73.23%	76.76%	16.41
	√			86.57%	86.57%	84.96%	69.23
	√	√		88.41%	88.41%	90.37%	53.05
	√		√	87.76%	87.26%	87.51%	68.85
	√	√	√	95.24%	89.02%	92.13%	52.15

5. CONCLUSION

In the quest for a digitized, automated, and intelligent approach to power grid planning, the transformation of the power grid into a digital counterpart is imperative. This transformation necessitates the creation of a "digital twin" for physical power equipment, facilitating precise classification and management. Our research introduces a pioneering deep learning-based method for the object detection of power equipment, leveraging staged training on dual-scale datasets to accommodate the diverse scales of power equipment. This methodology divides the dataset into large-scale and small-scale categories, with the model trained on each dataset to address the challenges of scale disparity, complex data labeling, and the nuanced learning requirements of smaller objects.

We meticulously constructed a dataset specific to power equipment identification, encapsulating a broad spectrum of scales, angles, and lighting conditions, comprising four distinct types of equipment: utility poles, transformers, insulators, and crossbars, all meticulously labeled. Through strategic sample selection and sophisticated data augmentation techniques, we have achieved an optimal balance between model generalization and applicability to specific scenarios. Our single-stage object detection model, tailored for data acquisition devices such as drones, sets a new standard in recognition accuracy and processing speed while requiring significantly fewer parameters than existing solutions.

The algorithm presented in this paper not only achieves an exemplary mAP of 92.13% and a detection speed of 52.15 fps but also demonstrates robust performance across a variety of natural conditions, including occlusion, shadow, and diverse weather phenomena. It adeptly handles the complexities of multiple scales, angles, and rotations in image acquisition, showcasing remarkable adaptability to the dynamic and intricate environments encountered in practical inspection scenarios. Notably superior to other deep learning-based object detection algorithms, our proposed method meets the stringent requirements for low-latency, multi-angle, and multi-target recognition and positioning of power equipment in the domain of intelligent grid line identification.

This research not only marks a significant advancement in the field of power equipment recognition but also lays the groundwork for the digitization and intelligent management of power grids. By striking a balance between accuracy, efficiency, and model compactness, our algorithm paves the way for the seamless integration of digital simulations and data management tasks. This contributes substantially to the realization of an intelligent power grid planning system, emphasizing our method's potential to revolutionize energy management practices and underscore its vital role in promoting sustainable and efficient power grid operations. Through this work, we envision fostering a future where digital twin technologies are at the forefront of innovative power grid planning and management, driving progress towards a more sustainable, efficient, and intelligent energy infrastructure.

ACKNOWLEDGEMENTS

This work was supported by Science and Technology Project of Hebei Electric Power Company "Research on Key Technologies for Planning of Digital Active Distribution Network in Xiong' an New Area" (Grant No.: SGHEJY00GHJS2000103).

REFERENCES

[1] Lu, Z., Deng, Y., Jiang, L., Li, J., Liu, Z., Liang, H., Zhang, T. (2022). Research on the influence of large-scale data center construction on power grid planning

under the background of east digital west computing. Journal of Global Energy Interconnection, 5(6): 552-562.
[2] Bai, H., Zhou, C., Yuan, Z., Lei, J. (2020). Prospect and thinking of digital power grid based on digital twin. Southern Power System Technology, 14(8): 18-24. https://doi.org/10.13648/j.cnki.issn1674-0629.2020.08.003
[3] Cheng, Z., Zhu, X., Li, J. (2022). Multi-criteria fusion decisionmaking method for power grid planning investment based on improved ELECTRE method. China Electric Power, 55(11): 59-65. https://doi.org/10.11930/j.issn.1004-9649.202201086
[4] Cheng, X., Song, C., Shi, J. (2021). A survey of generic object detection methods based on deep learning. Acta Electronica Sinica, 49(7): 1428-1438. https://doi.org/10.12263/DZXB.20200570
[5] Jiang, A., Yan, N., Wang, F., Huang, H., Zhu, H., Wei, B. (2019). Visible image recognition of power transformer equipment based on mask R-CNN. In 2019 IEEE Sustainable Power and Energy Conference (iSPEC), Beijing, China, pp. 657-661. https://doi.org/10.1109/iSPEC48194.2019.8975213
[6] He, K., Gkioxari, G., Dollar, P., Girshick, R. (2017). Mask R-CNN. In 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, pp. 2980-2988. https://doi.org/10.1109/ICCV.2017.322
[7] Ren, S., He, K., Girshick, R., Sun, J. (2016). Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6): 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031
[8] Lin, G., Wang, B., Peng, H., Wang, X., Chen, S., Zhang, L. (2019). Multi-target detection and location of transmission line inspection image based on improved faster-RCNN. Electric Power Automation Equipment, 39(5): 213-218.
[9] Wan, J., Wu, G., Guan, M., Wu, K., Gao, A., Shi, K. (2021). Intelligent recognition method for transformer small comonents based on RetinaNet. Power System Protection and Control, 12: 166-173.
[10] Chen, O., Qin, L., Yu, C. (2022). Improved rfbnet for power equipment infrared image recognition. Information Technology and Informatization, 2022(2): 108-111. https://doi.org/10.3969/j.issn.1672-9528.2022.02.028
[11] Liu, S., Huang, D. (2018). Receptive field block net for accurate and fast object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, pp. 404–419. https://doi.org/10.48550/arXiv.1711.07767
[12] Xiong, K., Fan, S., Wu, J. (2022). Identification method of substation power equipment in foggy weathers based on improved yolov4. Radio Engineering, 2022: 1504-1512.
[13] Girshick, R., Donahue, J., Darrell, T., Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, pp. 580-587. https://doi.org/10.1109/CVPR.2014.81
[14] Girshick, R. (2015). Fast r-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, pp. 1440-1448. https://doi.org/10.1109/ICCV.2015.169
[15] He, K., Zhang, X., Ren, S., Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(9): 1904-1916. https://doi.org/10.1109/TPAMI.2015.2389824
[16] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C. (2016). Ssd: Single shot multibox detector. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, Part I 14, pp. 21-37. https://doi.org/10.1007/978-3-319-46448-0_2
[17] Redmon, J., Farhadi, A. (2018). Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767. https://doi.org/10.48550/arXiv.1804.02767
[18] Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollar, P. (2020). Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(2): 318-327. https://doi.org/10.1109/TPAMI.2018.2858826
[19] Tan, M., Pang, R., Le, Q.V. (2020). Efficientdet: Scalable and efficient object detection. in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, pp. 10778-10787. https://doi.org/10.1109/CVPR42600.2020.01079
[20] Zhou, X., Wang, D., Krähenbühl, P. (2019). Objects as points. arXiv preprint arXiv:1904.07850. https://doi.org/10.48550/arXiv.1904.07850
[21] Law, H., Deng, J. (2018). Cornernet: Detecting objects as paired keypoints. In European Conference on Computer Vision, 8: 765–781. https://doi.org/10.48550/arXiv.1808.01244
[22] Shelhamer, E., Long, J., Darrell, T. (2017). Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(4): 640-651. https://doi.org/10.1109/TPAMI.2016.2572683
[23] Tian, Z., Shen, C., Chen, H., He, T. (1904). FCOS: Fully convolutional one-stage object detection. arXiv 2019. arXiv preprint arXiv:1904.01355.
[24] Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J. (2021). Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430. https://doi.org/10.48550/arXiv.2107.08430
[25] Redmon, J., Divvala, S., Girshick, R., Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, pp. 779-788. https://doi.org/10.1109/CVPR.2016.91
[26] Simonyan, K., Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. https://doi.org/10.48550/arXiv.1409.1556
[27] Tan, M., Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In International Conference on Machine Learning, California, pp. 6105-6114.
[28] Zoph, B., Cubuk, E.D., Ghiasi, G., Lin, T.Y., Shlens, J., Le, Q.V. (2020). Learning data augmentation strategies for object detection. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, Part XXVII 16, pp. 566-583. https://doi.org/10.1007/978-3-030-58583-9_34