IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 16, 2021, accepted May 2, 2021, date of publication May 11, 2021, date of current version May 24, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3079172

Detection of Pin Defects in Aerial Images Based on Cascaded Convolutional Neural Network

YEWEI XIAO[1,2], **ZHIQIANG LI**[1,2], **DONGBO ZHANG**[1,3], **AND LIANWEI TENG**[1,2]

[1] School of Information Engineering, Xiangtan University, Xiangtan 411105, China
[2] Key Laboratory of Intelligent Computing and Information Processing of Ministry of Education, Xiangtan University, Xiangtan 411105, China
[3] National Engineering Laboratory of Robot Vision Perception and Control Technology, Xiangtan University, Xiangtan 411105, China

Corresponding author: Zhiqiang Li (596210498@qq.com)

This work was supported in part by the Joint Fund for Regional Innovation and Development of NSFC under Grant U19A2083, and in part by the Science and Technology Research and Major Achievements Transformation Project of Strategic Emerging Industries in Hunan Province under Grant 2019 GK4007.

ABSTRACT Pins are standard fasteners in power transmission lines, and the hidden dangers of pins falling off dramatically affects their safe operation. If a pin is missed, it is called pin defects in this paper. As the pin is a small target and has a complex background, traditional detection algorithms were used to identify pin defects from aerial images which suffer from poor accuracy and low efficiency. This paper proposed a target detection method based on cascaded convolutional neural networks. First, a small-scale shallow full convolutional neural network was used to obtain the region of interest; then, a deeper convolutional neural network conducted target classification and positioning on the obtained region of interest. Next, a nonlinear multilayer perceptron was introduced, the convolution kernel was decomposed, and the multi-scale feature maps were fused. At this point, an angle variable was added to the classification cross-entropy loss function. Multi-task learning and offline hard sample mining strategies were used in the training phase. The proposed method was tested on a self-built pin dataset and the remote sensing image RSOD dataset, and the experimental results proved its effectiveness. Our method can accurately identify pin defects in aerial images, thereby solving the engineering application problem of pin defect detection in transmission lines.

INDEX TERMS Pin defect, aerial image, cascaded convolutional neural network, nonlinear multilayer perceptron, hard sample mining.

I. INTRODUCTION

Power fittings play an indispensable role in stable, safe, and reliable power system operation [1], [2]. However, due to the harsh environment and the influence of the power system load, power fittings are prone to missing pins, which will cause power system malfunction. Therefore, pin defect detection is of great significance for maintaining the safe operation of the power system.

In recent years, drone inspections have been widely used in the daily inspections of transmission lines, reducing the workload of grid operation and maintenance personnel in regards to climbing inspections, and thus also reducing the risk associated with such work [3]. Drone inspections can efficiently and accurately determine the fault condition of the equipment. Despite improvements in the power system, the massive increase in the amount of power transmission equipment, and the rapid growth of aerial image applications, the traditional manual identification and detection efficiency are still low [4]. Therefore, pin defect recognition faces severe challenges [5].

At present, most image inspection research is focused on large targets, such as pressure-equalizing rings, hanging plates, and insulators, and is rarely focused on pin-level defect recognition (i.e., small targets). In the traditional image inspection method, transmission line defect detection is based on manually designed features. For example, in the image preprocessing stage, operations such as image enhancement are often used. When realizing the defect detection of critical components of transmission lines, it is common to combine features such as color space, moment invariants, and Haar-like. Adaboost and other traditional algorithms are often used for defect detection and identification [6]–[8]. For example, Jin *et al.* classified the surface condition of the insulator

The associate editor coordinating the review of this manuscript and approving it for publication was Le Hoang Son.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

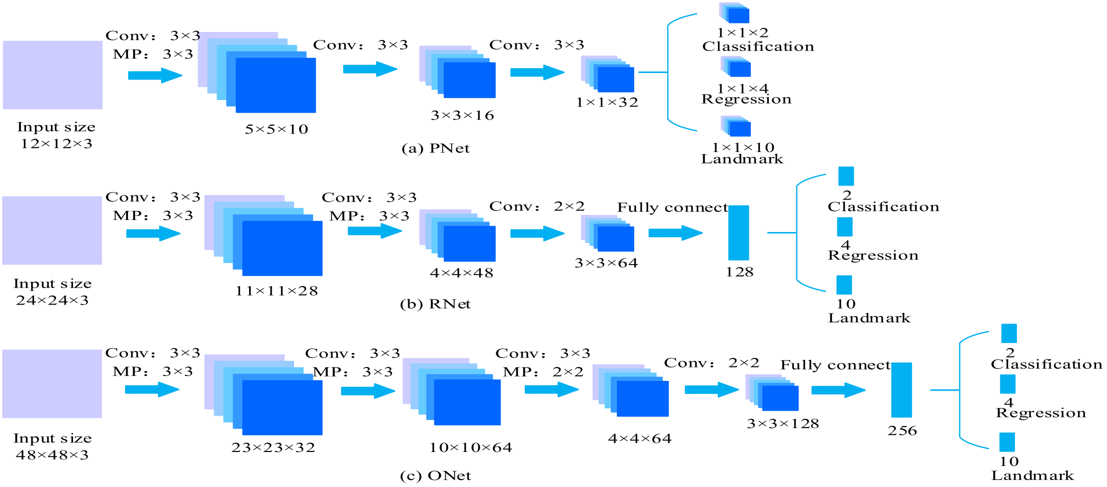


FIGURE 1. Structure of MTCNN.

by fusing the principal component analysis method with the different color features of the color space; meanwhile, literature [9] proposed the idea of two-step detection, which achieves a higher recognition rate, but has a slower detection speed. Since manually designing features is cumbersome and leads to low robustness, some scholars have proposed defect detection based on deep learning algorithms. For instance, literature [10] used a convolutional neural network to extract image sample features to identify self-explosion defects of insulators; furthermore, literature [11], [12] compared and analyzed the recognition performance of target detection algorithms such as Faster RCNN, SPPnet, and DPM, and investigated the impact on performance brought by key power components of transmission lines, such as vibration-dampers and insulators.

With the existing target detection methods, it is tricky to detect pin defects in large-scale aerial transmission line images with complex backgrounds. The main technical difficulties are as follows: (1) the images are blurred due to jitter and light intensity during collection by the drone; (2) the image background is complex, the detection target is small and occluded; (3) high-precision convolutional neural networks generally have a profound number of layers [13], [14], and the resulting calculation and storage costs are enormous; (4) a deeper network structure requires a lot of labeled data during training to ensure high detection accuracy, and the training is complex and inefficient; and (5) since existing target detection algorithms are designed for conventional images, the image size must be fixed during training and detection [15]–[18]. Thus, we cannot achieve versatility and scale adaptability in the target detection of aerial transmission line images.

According to the listed problems, this paper proposes a pin defect detection method for aerial images based on a multi-scale cascaded convolutional neural network. When processing large-scale images, a small-scale shallow full convolution is used to traverse aerial images and conduct the target search quickly. We then use a deeper convolution to achieve cascaded classification and precise positioning of the obtained candidate targets. In the shallow, fully convolutional neural network, aerial images of any scale can be input. Compared to a single-layer neural network, our network has faster detection speed and enhanced accuracy. Therefore, the proposed method presents advantages for application in mobile devices.

II. CASCADED CONVOLUTIONAL NEURAL NETWORK

The multi-task convolutional neural network (MTCNN) [19], as shown in Figure 1, includes three small convolutional neural networks of different scales: PNet, RNet, and ONet. PNet is a shallow fully convolutional network consisting of three convolutional layers and one pooling layer. RNet, ONet, and PNet are similar in structure, but the network structure of [RNet, ONet] is deeper, and thus the candidate targets can be classified and located more accurately.

In the direct use of the MTCNN in pin defect detection and detection experiments on a self-built pin dataset, we found that there is still a large number of false detections and missed detection targets; moreover, the detection accuracy is not high. The main reasons for this are as follows: (1) there are few images of missing pins in the dataset, resulting in an imbalance between the standard samples and defective samples in the training set, which leads to an insufficient classification effect; and (2) pins are so small that it is not easy to distinguish aerial images from the background image (we call these pins targets hard samples). Overall, due to the unbalanced sample categories and existence of hard samples, the MTCNN faces great challenges in pin defect identification and detection.

III. IMPROVED METHOD

We improve the original MTCNN in the following four aspects: (1) after decomposing the convolution kernel, we add a nonlinear multilayer perceptron after the partial convolution

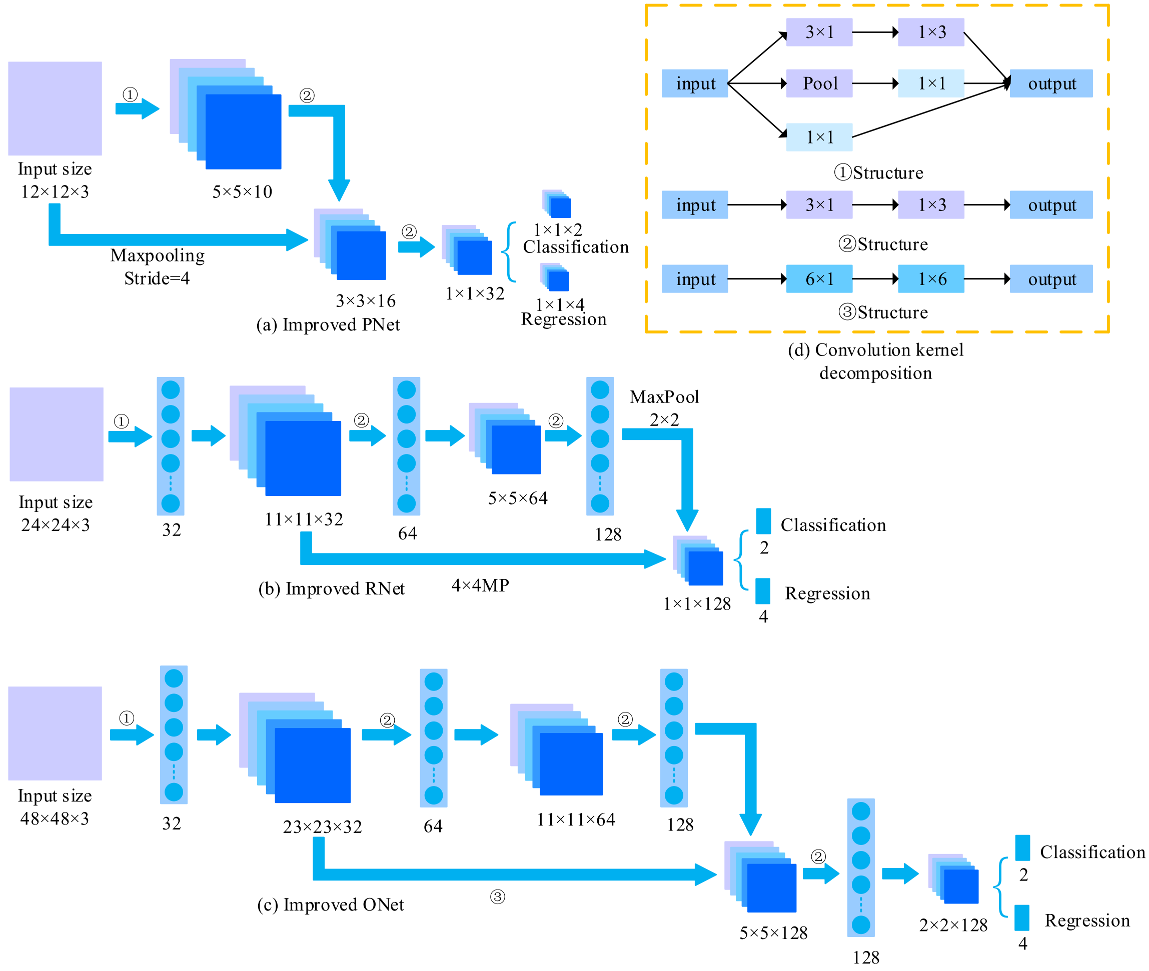
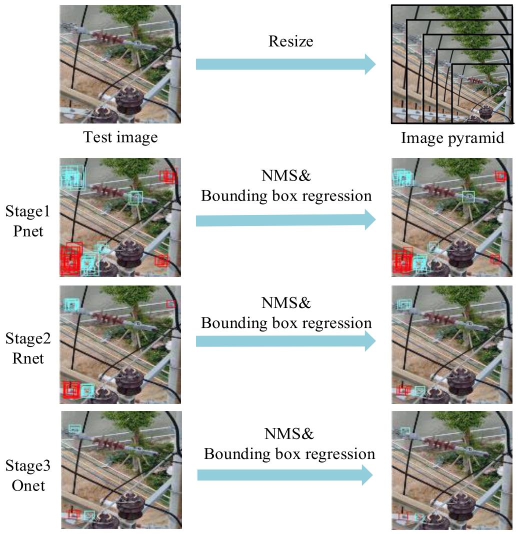


FIGURE 2. Improved MTCNN.

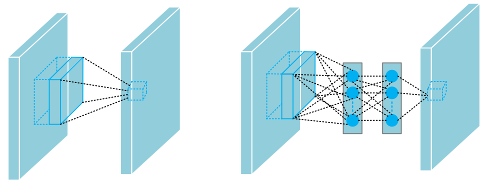
layer, and simultaneously remove the fully connected layer; (2) we fuse multi-scale feature maps; (3) we add an angle variable to the classification cross-entropy loss function; and (4) we adopt multi-task learning and offline hard sample mining strategies in the training process. The overall framework of the improved network is shown in Figure 2. Because PNet's role is to generate regions of interest, the image input to the network is small (12 × 12 pixels). The improved accuracy is not apparent after adding a nonlinear multilayer perceptron after the convolutional layer; instead, it reduces the speed and efficiency of the model. Thus only the fusion of the feature map and the decomposition of the convolution kernel are performed in PNet.

The detection process is a form of cascade detection, as shown in Figure 3. First, we pyramid the input image, and use the sliding window method in the candidate network PNet to process the candidate frame. At the same time, we use the non-maximum suppression algorithm NMS in the detection process to reduce the generation of overlapping boxes. We then map the candidate frame obtained in each image to the original image to obtain the target slice, and then use the optimized network RNet to classify the candidate target and return the boundary. Finally, the output network ONet is used to classify and perform boundary regression on the candidate frames that reach the specified threshold.



Test image
Resize
Image pyramid
Stage1 Pnet
NMS& Bounding box regression
Stage2 Rnet
NMS& Bounding box regression
Stage3 Onet
NMS& Bounding box regression

FIGURE 3. Flowchart of cascade detection on test image.



(a) Linear convolution layer (b) Mlpconv layer

FIGURE 4. **Improved convolutional layer.**

A. IMPROVED CONVOLUTIONAL LAYER STRUCTURE

The simple network structure of the original MTCNN leads to weak feature information extracted by the convolutional layer. Besides, there are many fully connected layer parameters used for classification in RNet and ONet, which easily cause overfitting, and thus cause the model's generalizability to be weak. To improve the overall performance of the model, we need to optimize the network structure. In the traditional convolution process, data processing is a generalized linear transformation, and the ability to abstract data features is weak. Here, we refer to the work of [20]. We add a large number of 1 × 1 convolutions to the traditional convolutional neural network structure, as shown in Figure 4 below.

Traditional convolutional neural networks generally consist of linear convolutional layers, pooling layers, and fully connected layers. With the help of a linear filter and a nonlinear activation function, the convolution layer is connected to realize the generation of the feature map. Taking ReLU as an example, the calculation of the feature map is as follows.

$$f_{i,j,k} = \max\left(w_k^T x_{i,j}, 0\right) \tag{1}$$

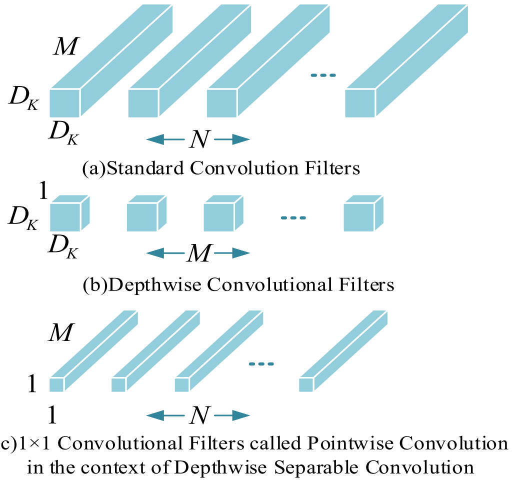
In this formula, (i, j) represents the position index of the image pixel, $x_{i,j}$ is an input image block at position (i, j), and k represents the index of the feature map we want to extract.

The traditional convolution is not sufficient for highly nonlinear feature extraction. The Mlpconv structure proposed in [20] can effectively solve this problem. The Mlpconv layer can be employed since each convolutional local receptive field contains a miniature multilayer network. With this multilayer Mlp micro network's help, more complicated calculations can be implemented for the neurons in each local receptive field. The calculation of each feature map of the Mlpconv layer is as follows:

$$\begin{aligned} f^1_{i,j,k_1} &= \max(w^{1\,T}_{k_1} x_{i,j} + b_{k_1}, 0). \\ &\vdots \\ f^n_{i,j,k_n} &= \max(w^{n\,T}_{k_n} f^{n-1}_{i,j} + b_{k_n}, 0). \end{aligned} \tag{2}$$

Here, n represents the n-th layer, and b_{kn} represents the offset.

In this way, the features with stronger generalizability and higher abstraction can be obtained, and the dimensionality of the data can be reduced. Moreover, replacing the fully connected layer in the traditional CNN with global pooling can reduce the number of network parameters as well as the possibility of overfitting.



(a)Standard Convolution Filters

(b)Depthwise Convolutional Filters

(c)1×1 Convolutional Filters called Pointwise Convolution in the context of Depthwise Separable Convolution

FIGURE 5. **Depth separable convolution.**

After modifying the network structure, the accuracy of target detection is improved, and the complexity of the network is increased. Generally speaking, this will reduce the detection speed of the model. To solve this problem to a certain extent, literature [21] proposed the depth separable convolution. The MobileNet was established based on the deep separable convolution, which is a type of decomposed convolution. As shown in Figure 5, the standard convolution is resolved into a deep convolution; note that a 1 × 1 convolution is also called a point-wise convolution. For MobileNet, the deep convolution applies a convolution kernel on each channel, and then the 1 × 1 point-wise convolution combines the output and the deep convolution. The standard convolution combines the input and the convolution kernel into a new output in one step. The separable depth convolution divides this output into two layers: one for filtering, and the other for combination. This decomposition process significantly reduces the amount of calculation in the model.

In the cascaded MTCNN, the convolution operation occupies most of the network's computing resources. Hence, we use the idea of depth separable convolution presented in literature [21] to decompose the convolution kernel in the network, thereby significantly reducing the amount of calculation. To further improve the network's detection ability, the first-layer 3 × 3 convolution operation of PNet, RNet, and ONet aggregates and reconvolves the features on multiple scales. The improved structure of the convolution kernel is shown in Figure 2(d).

B. FEATURE MAP FUSION

To improve pin defect recognition performance, we added feature map fusion [22] to the cascaded MTCNN. Literature [22] combined the detection results of different layers to improve detection performance with feature map fusion, predicted multi-scale features separately, and then fused the prediction results. In the present paper, we follow this idea to improve the cascaded MTCNN. First, the 12 × 12 pixel detection window in PNet is subjected to three-layer convolution

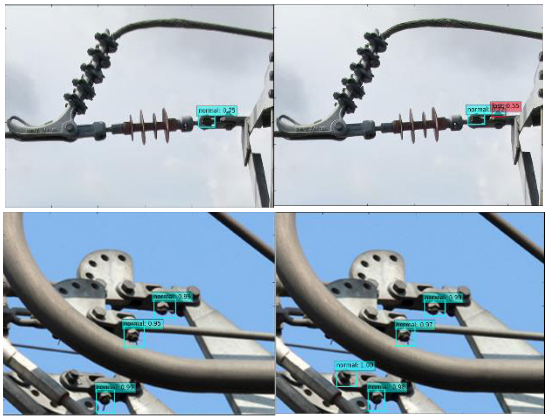


FIGURE 6. Different detection effects of the original model structure and the model with feature map fusion.

to calculate the loss. Then, we fuse feature maps of the first and third layers of PNet. Similarly, feature map fusion is performed on the second and third layers of RNet and ONet.

Figure 6 compares the detection effects between the original model structure and the model with feature map fusion. There are two sets of images. The first set contains two detection effect images to compare the detection effect of blocked pin defect targets. The second set contains two detection effect images to compare the detection effect for pins whose characteristics are not prominent. It can be seen that the detection effect with feature map fusion included in the network structure is more accurate than that without feature map fusion.

C. IMPROVED LOSS FUNCTION

In the MTCNN, the cross-entropy loss function is used for classification tasks, and is expressed as

$$L_i^{\text{det}} = -\left(y_i^{\text{det}} In\left(p_i\right) + \left(1 - y_i^{\text{det}}\right)\left(1 - In\left(p_i\right)\right)\right) \tag{3}$$

Here, L_i^{det} represents the loss, and y_i^{det} represents the network output. $p_i = exp(W_i^T x + b_j)/\sum_j exp(W_i^T + b_j)$ denotes the softmax of network output, where $W_i^T x + b_j$ is the output of the i-th output neuron, and W_i is the weight vector of the i-th neuron. Since the function uses a plane as a classification surface only, the distance between classes is too small, resulting in poor classification of samples near the classification plane. A useful loss function has a sufficiently small intra-class distance and a considerable inter-class distance. Thus, the classification plane of the loss function of the MTCNN is

$$\left(W_1 - W_2\right) x + b_1 + b_2 = 0 \tag{4}$$

We transform this into a formula about the angle between W_i and x with the help of an angle variable:

$$\left\| W_i^T \right\| \cdot \|x\| \cos\left(\theta_i\right) + b_i \tag{5}$$

Here, θ_i is the angle between W_i and x. Letting $||W_1|| = ||W_2|| = 1$, and $b_1 = b_2{=}0$ be the form of the decision boundary, we have

$$\|x\| \left(\cos\left(\theta_1\right) - \cos\left(\theta_2\right)\right) = 0 \tag{6}$$

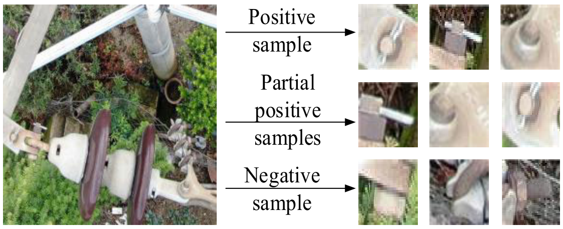


FIGURE 7. Samples for training.

On this basis, the restrictions for the loss function are further strengthened, that is, theangle variables are introduced into the classification by the loss function to better classify the detection targets. The loss function is

$$L = \frac{1}{N} \sum_i -In \left[\exp\left(\|x_i\| \cos\left(3\theta_{yi,i}\right) - 0.4\right) \middle/ \exp\left(\|x_i\| \cos\left(3\theta_{yi,i}\right)\right) + \sum_{j \neq y_i} \|x_i\| \cos\left(3\theta_{j,i}\right) \right] \tag{7}$$

Here, L represents the loss function, and θ is the angle between the classification surface and W. The decision boundary is

$$\|x\| \left(\cos\left(3\theta_1\right) - \cos\left(\theta_2\right)\right) = 0 \tag{8}$$

and

$$\|x\| \left(\cos\left(\theta_1\right) - \cos\left(3\theta_2\right)\right) = 0 \tag{9}$$

After adjustment, the distance between classes is effectively increased, the indexability of the decision-making area is improved, and the distribution of the angle $[\theta]$ within classes is compressed.

D. SAMPLE TRAINING

1) TRAINING SAMPLE SELECTION

To expand the number and diversity of training data, the training samples are divided into fourcategories: positive samples, negative samples, partial positive samples, and hard samples. Positive samples and negative samples can improve the model's classification loss function, while positive samples and partial positive samples can improve the model's boundary regression loss function. We divide the samples into positive samples, negative samples, and partial positive samples with the help of the intersection over union (*IOU*). The *IOU* is a concept used to evaluate positioning accuracy in target detection tasks. It is specifically defined as the ratio of the intersection of the prediction frame and the label frame area to the union. The formula is as follows:

$$IOU = \frac{area\left(gt \cap dt\right)}{area\left(gt \cup dt\right)} \tag{10}$$

where *gt* represents the target label box, and *dt* represents the target prediction box.

We use the sliding window method to select training samples, pyramid the training images, and then use a sliding

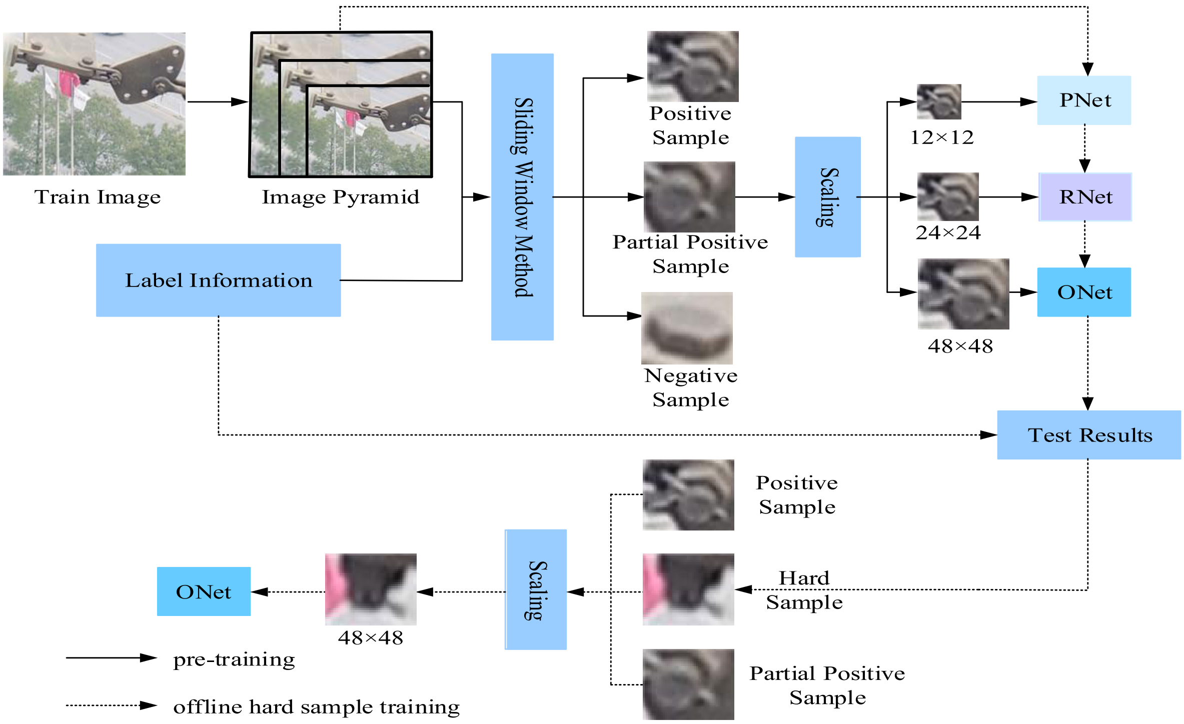


FIGURE 8. Flowchart of training.

window of a prescribed size to select a region of the image. As shown in Figure 7, the *IOU* of the selected area and the label box are calculated at the same time. An area with an *IOU* greater than 0.7 is recorded as a positive sample, an area with an *IOU* less than 0.3 is recorded as a negative sample, and an area with an *IOU* between 0.5–0.7 is recorded as a partial positive sample. Because the features extracted by the convolutional neural network have a certain degree of abstraction, it is difficult to judge and select hard samples.

2) TRAINING STRATEGY

To maximize the network's generalizability under the condition of limited labeled data, the training process is divided into two steps: pre-training, and offline hard sample training. The whole process is shown in Figure 8.

For pre-training, we adopt the strategy of online hard sample mining (OHEM) [23]. The propagation loss calculated by each batch of data is sorted, and the samples with the most considerable propagation loss are divided into hard samples according to a certain proportion. When performing back-propagation, only the loss of hard samples is used to update the weights in the neural network model. When using the pre-trained MTCNN, a large number of false and missed targets can be obtained. These targets are hard samples. The OHEM strategy ignores the gradient of some easily distinguishable samples when the convolutional neural network propagates backward. Moreover, since most of the selected samples are simple samples, it is difficult for the model to distinguish hard samples efficiently. This is the main factor causing false detections and missed detections. Adding hard samples to the training data can improve the false detection rate and missed detection rate. Therefore, in the offline hard sample training stage, the positive samples, negative samples, and partial positive samples are all scaled to 12 × 12 pixels which corresponds to PNet, 24 × 24 pixels which corresponds to RNet; meanwhile, PNet and the improved RNet are trained separately. Finally, the acquired hard samples, positive samples, and partial positive samples are scaled to 48 × 48 pixels for retraining on the improved ONet.

To verify the effectiveness of the added offline hard sample training method in improving detection accuracy, we trained the network on 100,000 iterations of self-built pin images. A test was performed every 4000 iterations. The training results are shown in Figure 9. It can be seen from Figure 9 that the original training method has 89.05% and 92.03% detection accuracy for missing pins and normal pins, i.e., pins without defects, respectively. In contrast, the improved training method has 90.17% and 93.85% detection accuracy for missing and normal pins. Experiments verify that the training method introducing offline hard samples yields an accuracy improvement of more than 1% for both types of pin detection.

IV. EXPERIMENT AND ANALYSIS

A. DATASET INTRODUCTION

Since there is currently no public normal pin database, we built our own pin dataset. To ensure the diversity of data, the data came from three places: (1) aerial images of high-voltage transmission lines; (2) aerial images of low-voltage

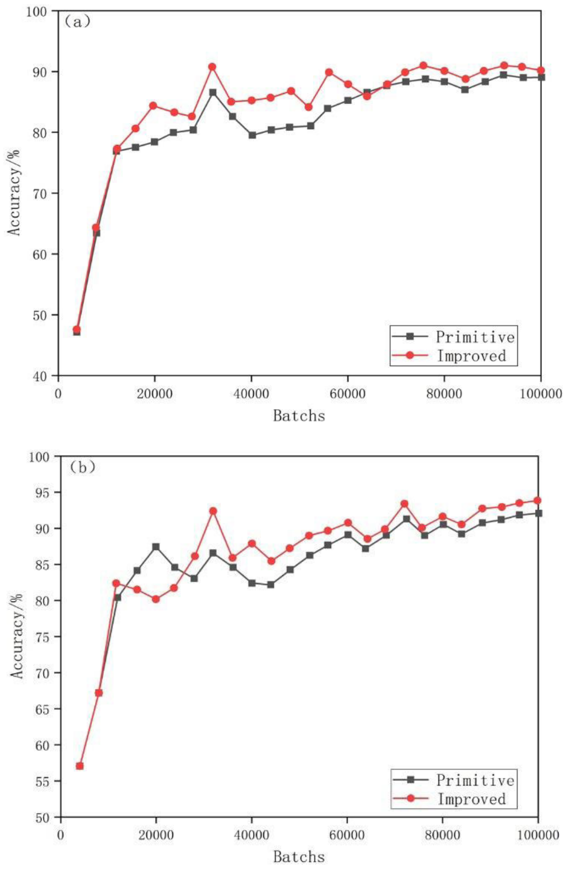


FIGURE 9. Accuracy changes under different training methods: (a) change in pin defect detection accuracy; (b) change in normal pin detection accuracy.

transmission lines; and (3) images obtained by Internet search. Including the three places, the images totaled 1500. 6000 aerial images were obtained after data augmentation processing. To ensure balance between normal pin targets and pin defect targets in the training samples, of the 70% of images selected for the aerial image training set, there were 11026 normal pin targets and 10654 pin defect targets; these comprised a total of 21680 positive samples. Using the training sample selection method in section 2.4.1, we selected 21680 positive samples and 65040 negative samples. The remaining 30% of aerial images were used in the test set to verify the network's generalizability and practicality. Sample images from the dataset are shown in Figure 10.

B. EXPERIMENTAL ENVIRONMENT AND ALGORITHM EVALUATION INDEX

The computer used in the experiment is configured with GTX1080Ti GPU, 11GB video memory, i7-8700 CPU, and the environment are Ubuntu16.04 and TensorFlow. The training set's batch size is 64, which depends on the GPU memory to make the algorithm converge faster. To objectively evaluate the performance of the model algorithm, this paper uses the Precision-Recall curve (P-R curve) and Average Precision (AP) and mean Average Precision (mAP) to conduct a comprehensive test on this algorithm.

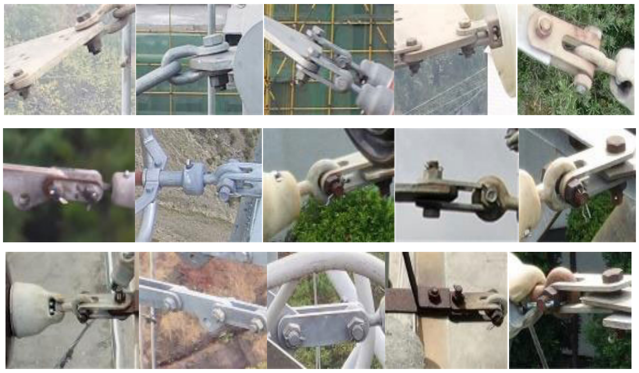


FIGURE 10. Part of the data set.

For the target detection task, to reflect the accuracy of predicting the target position, it is necessary to consider the *IOU* of the prediction frame and the real target frame when calculating the P-R curve. In this experiment, the *IOU* is set to 0.5.

C. EXPERIMENTAL RESULTS AND ANALYSIS

1) EXPERIMENT 1: PERFORMANCE EXPERIMENT OF THE IMPROVED METHOD

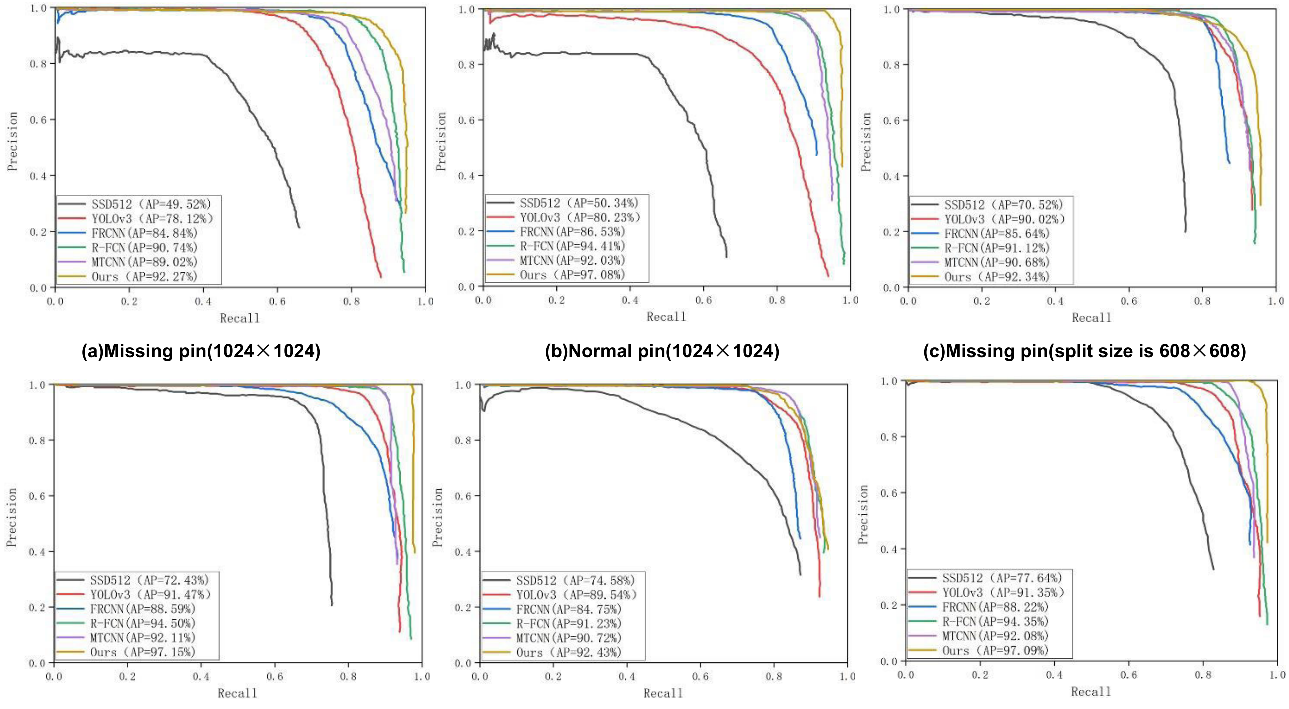
Generally speaking, the change of neural network structure has different positive and negative effects on the algorithm model's accuracy and speed. Considering the damage of pin defects to the power system, we can sacrifice part of the detection speed to improve the recognition accuracy. To verify the effectivenessof the several methods proposed in this article, several sets of experiments with different improved methods are set up. The experimental results are shown in Table 1.

It can be seen from method 1 and 2. 5ms improve the detection speed compared with the original MTCNN algorithm, the detection average precision of missing pins is increased by 1.1%, and the detection average precision of standard pins is increased 1.81%, and the performance improvement is more obvious. It can be seen from method 2 and 3 that the detection speed has increased by 3ms, and the detection average precision has increased by 0.84% and 1.02% for missing and normal pins, respectively. Considering the high accuracy requirements in this application scenario, it is feasible to increase the accuracy by nearly 1% at the expense of 5ms detection speed. It can be seen from method 3 and 4 that when the detection speed has not changed, the detection accuracy has increased by 1.12% and 1.82% for the missing and normal pins, respectively, which shows that the training strategy is applied to the sample category imbalanced data set Great advantage. It can be seen from methods 4 and 5 that the detection average precision rate has increased by 0.16% and 0.4%, respectively, on missing and standard pins. Since the algorithm only performs forward propagation during testing, the loss function's improvement only affects the training time and does not affect the test time.

It can be seen from the experimental data that the detection average precision of the several improved methods proposed in this paper has been improved based on ensuring

TABLE 1. Experimental results.

Improved method	Improved convolutional layer structure	Feature map fusion	Improved training strategy	Improved loss function	Detection time (ms)	Missing pin (AP)%	Normal pin (AP)%	(mAP)%
Method 1	×	×	×	×	40	89.05	92.03	90.54
Method 2	√	×	×	×	38	90.15	93.84	92.00
Method 3	√	√	×	×	43	90.99	94.86	92.93
Method 4	√	√	√	×	43	92.11	96.68	94.40
Method 5	√	√	√	√	43	92.27	97.08	94.68



(a)Missing pin(1024×1024)

(b)Normal pin(1024×1024)

(c)Missing pin(split size is 608×608)

(d)Normal pin(split size is 608×608)

(e)Missing pin(split size is 512×512)

(f)Normal pin(split size is 512×512)

FIGURE 11. Precision–recall curves of different networks on the self-built pin datasets.

the detection speed. The missing pins category increased by 3.22%, and the normal pins category increased by 5.05%. In terms of performance improvement, the effect of normal pins is better than missing pins. The key reason is that there are separate nuts on the fixed parts in the power transmission line. Compared with the nuts with missing pins, there is only a difference in whether there is a pinhole, which results in weak pin defect characteristics and low recognition. However, on the whole, it reflects the effectiveness of the proposed method.

2) EXPERIMENT 2: PERFORMANCE COMPARISON OF DIFFERENT ALGORITHMS

In order to verify the effectiveness of the improved algorithm model and training strategy in this paper, the same training and testing were performed on the self-built data set. Compare 4 mainstream algorithms for target detection: SSD512 [22], YOLOv3 [24], Faster RCNN, R-FCN [25] and the original MTCNN algorithm. The backbone network of SSD512 is VGG16 [13], the backbone network of YOLOv3 is Darknet53, the backbone network of Faster RCNN is ResNet50 [14],and the backbone network of R-FCN is ResNet101. In order to eliminate the influence of image size on the detection effect of SSD512 and YOLOv3 algorithms, for the self-built pin data set, two segmentation methods of 512 × 512 pixels and 608 × 608 pixels were adopted for training and detection. The specific experimental results are shown in Table 2 and Figure 11.

From Figure 11, we can clearly observe that some networks have better detection effects for normal pins than for missing pins. Among the six networks, SSD512 has the worst performance. The main reasons for this are as follows: (1) the

TABLE 2. Performance comparison of different networks.

Size Method	1024×1024			608×608			512×512		
	Missing pin (AP)%	Normal pin (AP)%	(mAP)%	Missing pin (AP)%	Normal pin (AP)%	(mAP)%	Missing pin (AP) %	Normal pin (AP)%	(mAP)%
SSD512	49.52	50.34	49.93	70.52	72.43	71.48	74.58	77.64	76.11
YOLOv3	78.12	80.23	79.18	90.02	91.47	90.75	89.54	91.35	90.45
Faster RCNN	84.84	86.53	85.69	85.64	88.59	87.11	84.75	88.22	86.48
R-FCN	90.74	94.41	92.58	91.12	94.50	92.81	91.23	94.35	92.79
MTCNN	89.02	92.03	90.54	90.68	92.11	91.40	90.72	92.08	91.40
Ours	92.27	97.08	94.68	92.34	97.15	94.75	92.43	97.09	94.76

images in the self-built pin dataset are 1024 × 1024, and when training SSD512, the image is severely scaled, which dramatically affects its training and detection performance; and (2) the backbone network of SSD512 is VGG16, which has poor generalization. YOLOv3's performance is also low, and its input image size is 608 × 608. After segmenting the images in the self-built pin dataset, the detection effects of SSD512 and YOLOv3 increase significantly, reaching accuracies of 70% and 80%, respectively; in contrast, the detection accuracy of other networks does not change significantly after this operation. The detection effect of R-FCN is higher than those of SSD512 and YOLOv3. The main reason for this is that the single-stage network sacrifices detection accuracy when improving the detection speed, especially when detecting small targets. Moreover, R-FCN uses the idea of region suggestion, a method of selecting samples, which yields better detection accuracy. The original MTCNN is superior to Faster RCNN in detection accuracy, showing that the cascaded convolutional neural network has inherent advantages in the multi-scale traversal and search method. After improving the MTCNN, the detection effect becomes significantly better than that of R-FCN.

We further compare the original MTCNN to the improved MTCNN by presenting some of the detection images in Figure 12. It can be seen from the detection results that under certain circumstances, the original MTCNN has some missed detections and false detections, and thus it cannot be used in actual engineering applications. In contrast, the improved MTCNN has strong robustness.

3) EXPERIMENT 3: PERFORMANCE COMPARISON ON ANOTHER DATASET

Here we further verify the effectiveness and practicality of the proposed network by testing it on another dataset. We employ the public RSOD [26] dataset of remote sensing images to detect aircraft targets (image size ranges from 1024 × 768 to 1044 × 915 pixels). This dataset has similarities with the self-built pin dataset in image size and the presence of small targets. The networks used for comparison are the same as those in Experiment 2. In Figure 13 we draw the P–R curves of each network. It can be seen from this comparison that the performance of the improved MTCNN on RSOD is significantly better than that of other networks, thereby reflecting its strong robustness.

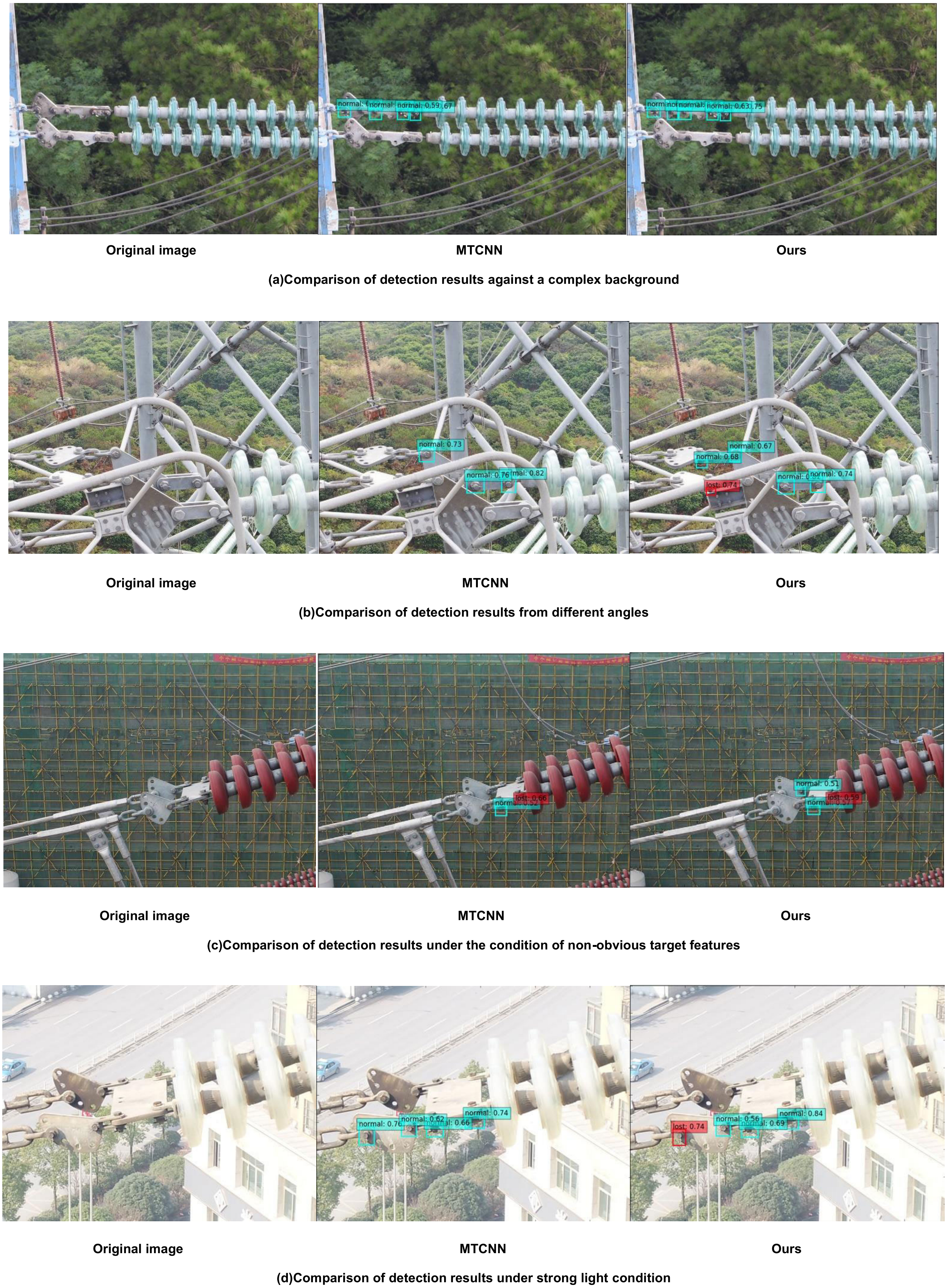
4) EXPERIMENT 4: ANALYSIS OF THE MINIMUM RESOLUTION OF IMAGES FOR SUFFICIENT DETECTION EFFECT

In actual situations, the quality of images obtained by UAVs varies due to external factors when patrolling the power grid. The higher the resolution, the higher the image quality, and vice versa. Therefore, it is necessary to verify how well the algorithm in this paper can detect the minimum resolution. The images comprising the self-built pin dataset have a spatial resolution of 0.2 m, and the UAV shooting distance is 5–10 m. To obtain images of different resolutions, we send an unmanned aerial vehicle to collect images with resolutions of 0.1 m and 0.08 m; again, the UAV shooting distance is 5–10 m. Then, we conduct detection experiments using the same set of networks. The experimental results are shown in Table 3.

It can be seen from the experimental results that the detection accuracy of all networks except SSD512 suffers greatly when the resolution is halved. SSD512's immunity is due to the fact that its performance is already very unsatisfactory under the resolution of 0.2 m (the specific reason for this is already discussed in Experiment 2). Under the resolution of 0.08 m, our proposed network obtains a detection accuracy greater than 80%. Hence, our network can ensure good detection of missing pins when the image quality varies, as in actual situations.

D. NETWORK SIZE AND TIME CONSUMPTION

At present, although ShuffleNet, MobileNet, and other networks have achieved good results, lightweight networks generally cannot achieve the accuracy of deep networks. Table 4 lists the network size and detection time of several target detection networks. Deep convolutional neural networks have reached hundreds of megabytes, but cascaded small convolutional neural networks have apparent advantages in



Original image MTCNN Ours

(a)Comparison of detection results against a complex background

Original image MTCNN Ours

(b)Comparison of detection results from different angles

Original image MTCNN Ours

(c)Comparison of detection results under the condition of non-obvious target features

Original image MTCNN Ours

(d)Comparison of detection results under strong light condition

FIGURE 12. Comparison of detection results in specified contexts.

TABLE 3. Detection performance under different spatial resolutions.

Spatial resolution	0.2 m			0.1 m			0.08 m		
	Missing pin	Normal pin	(mAP)	Missing pin	Normal pin	(mAP)	Missing pin	Normal pin	(mAP)
Method	(AP)%	(AP)%	%	(AP)%	(AP)%	%	(AP)%	(AP)%	%
SSD512	49.52	50.34	49.93	47.34	48.15	47.75	46.42	47.22	46.82
YOLOv3	78.12	80.23	79.18	68.00	70.18	69.09	65.89	67.52	66.71
Faster RCNN	84.84	88.78	86.81	73.25	77.58	75.42	71.75	75.34	73.55
R-FCN	90.74	94.41	92.58	80.67	84.30	82.49	77.56	82.28	79.92
MTCNN	89.02	92.03	90.54	79.13	82.32	80.73	77.07	80.29	78.68
Ours	92.27	97.08	94.68	82.37	86.12	84.25	80.15	84.10	82.13

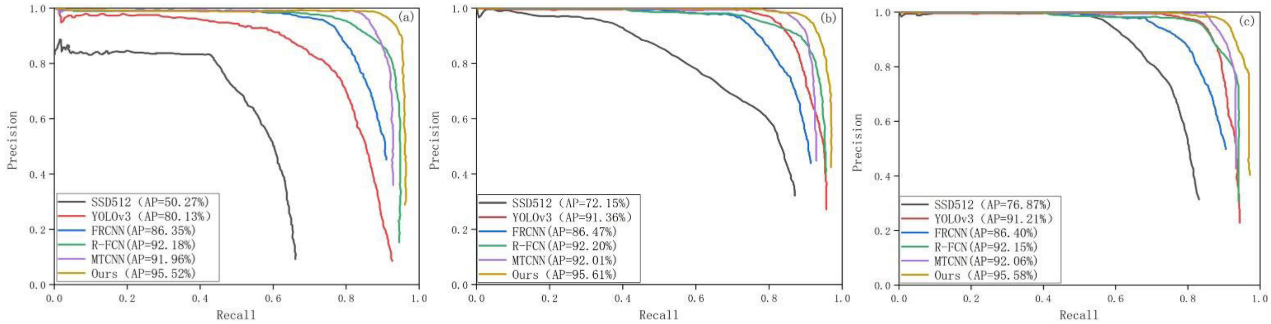


FIGURE 13. Precision-Recall of different algorithms in RSOD data sets: (a) Detection results of RSOD data sets; (b) Detection results of the 608 × 608 image split from RSOD data sets; (c) Detection results of the 512 × 512 image split from RSOD data sets.

TABLE 4. Comparison of network size and detection time.

Method	Size (M)	Detection time (ms)
SSD512	267	--
YOLOv3	236	51
Faster RCNN	208	1869
R-FCN	227	1973
MTCNN	3.9	40
Ours	3.2	43

Note: The size refers to the number of bytes saved in the model parameters.

being lightweight. This means that the network can process images more efficiently and be applied to mobile platforms. Because the convolutional layer has a significant impact on the network's running speed, merely increasing the number of convolutional layer parameters without correspondingly increasing the number of network layers has no obvious effect on accuracy improvement, and dramatically increases the number of parameters. Our proposed method shows significant improvement in performance, and at the same time reduces the number of parameters, which reduces hardware requirements to a certain extent.

In this comparison, the image is from the self-built pin dataset, and is 1024 × 1024. Because the detection accuracy of SSD512 is so low, it has no application value, and thus its detection time is omitted. Under the same hardware conditions, the high-efficiency YOLOv3 network is also at a disadvantage compared to the convolutional cascaded neural network in detection speed. When detecting images, the convolutional cascaded neural network method uses image pyramids to detect images of different scales. This method is time-consuming, but the detection speed is also close to real-time detection. However, the original MTCNN is slightly faster in detection speed than the proposed network. Considering the improved accuracy brought about by our proposed network and the particular scene of pin defect detection for power transmission lines, this small sacrifice in detection time is completely acceptable and our network remains feasible for practical engineering applications.

V. CONCLUSION

In this paper, we proposed a cascaded convolutional neural network for pin defect recognition and detection. Our network included convolution kernel decomposition, re-aggregation on PNet, feature map fusion of RNet and ONet, an improved convolution structure, and an improved training strategy. Comparative analysis of different mainstream target detection networks on our self-built pin dataset and the public remote sensing image dataset RSOD verified that our network could achieve better accuracy and support large-scale image target detection. Furthermore, our detection process used the pyramid multi-level detection method to solve the problem of images being of different scales.

The deep convolutional neural network's primary purpose is to perform multi-target detection. Although it is not totally precise that the deep convolutional neural network is compared with the cascaded convolutional neural network from the perspective of detection accuracy, detection speed, and number of parameters in this paper, it still can be seen that the latter is featuring a simpler model, less parameters, a faster speed and a better portability on mobile platforms when it comes to the target detection of specified tasks and fields. These advantages of the cascaded convolutional neural network should not be ignored.

REFERENCES

[1] Z. Xin and Y. Yingchun, "Effect of mesh method on strain numerical simulation for electric fittings," *Yunnan Electr. Power*, vol. 42, no. 1, pp. 71–73, 2014.

[2] L. Wang, K. Qian, M. Qian, Y. Sheng, and J. Hong, "Technical description and application of abrasion test methods for fittings of transmission lines," *Zhejiang Electr. Power*, vol. 35, no. 11, pp. 60–66, 2016.

[3] R. J. Liao, Y. Y. Wang, and H. Liu, "Research status of condition assessment method for power equipment," *High Voltage Eng.*, vol. 44, no. 11, pp. 3454–3464, 2018.

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

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

[6] L. Jin, S. Yan, and Y. Liu, "Vibration Damper recognition based on Harr-like feature and cascade Adaboost classifier," *J. Syst. Simul.*, vol. 24, no. 9, pp. 1806–1809, 2012.

[7] C. Wenming, W. Yaonan, Y. Feng, W. Xiru, and M. Siyi, "Research on obstacle recognition based on vision for deicing robot on high voltage transmission line," *Chin. J. Sci. Instrum.*, vol. 32, no. 9, pp. 2049–2056, 2011.

[8] L. Jin, D. Zhang, S. Duan, and S. Yao, "Contamination grades measurement of insulators based on image color feature fusion," *J. Tongji Univ.*, vol. 42, no. 10, pp. 1612–1617, 2014.

[9] L. Huaiyuan, "Study on visual identification method of obstacle on high voltage transmission line," Harbin Inst. Technol., Harbin, China, Tech. Rep., 2017, vol. 16, no. 3, pp. 25–36.

[10] Q. Chen, B. Yan, R. Ye, and X. Zhou, "Insulator detection and recognition of explosion fault based on convolutional neural networks," *J. Electron. Meas. Instrum.*, vol. 31, no. 6, pp. 942–953, 2017.

[11] T. Yong, J. Han, W. Wei, D. Jian, and X. Peng, "Research on part recognition and defect detection of trainsmission line in deep learning," *Electron. Meas. Technol.*, vol. 41, no. 6, pp. 60–65, 2018.

[12] W. Wang, B. Tian, Y. Liu, L. Liu, and J. Li, "Study on the electrical devices detection in UAV images based on region based convolutional neural networks," *J. Geo-Inf. Sci.*, vol. 19, no. 2, pp. 256–263, 2017.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[15] H. Tao, "Specific object detection and recognition in optical remote sensing image," Univ. Electron. Sci. Technol., Chengdu, China, Tech. Rep., 2018, vol. 17, no. 2, pp. 42–58.

[16] P. Xin, Y. L. Xu, and H. Tang, "Fast airplane detection based on multilayer feature fusion of fully convolutional networks," *Acta Optica Sinica*, vol. 38, no. 3, 2018, Art. no. 0315003.

[17] X. Han, Y. Zhong, and L. Zhang, "An efficient and robust integrated geospatial object detection framework for high spatial resolution remote sensing imagery," *Remote Sens.*, vol. 9, no. 7, p. 666, Jun. 2017, doi: 10.3390/rs9070666.

[18] M. Radovic, O. Adarkwa, and Q. Wang, "Object recognition in aerial images using convolutional neural networks," *J. Imag.*, vol. 3, no. 2, p. 21, Jun. 2017, doi: 10.3390/jimaging3020021.

[19] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016, doi: 10.1109/LSP.2016.2603342.

[20] M. Lin, Q. Chen, and S. Yan, "Network in network," 2013, *arXiv:1312.4400*. [Online]. Available: https://arxiv.org/abs/1312.4400

[21] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. 14th Eur. Conf. Comput. Vis.* Amsterdam, The Netherlands: Springer, 2016, pp. 21–37.

[22] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*. [Online]. Available: https://arxiv.org/abs/1704.04861

[23] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 761–769.

[24] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Dec. 2018, pp. 3–11.

[25] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2016, pp. 2–8.

[26] Y. Long, Y. Gong, Z. Xiao, and Q. Liu, "Accurate object localization in remote sensing images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2486–2498, May 2017, doi: 10.1109/TGRS.2016.2645610.

YEWEI XIAO received the B.S. and M.S. degrees from Xiangtan University, Xiangtan, China, in 2000 and 2004, respectively. He is currently an Associate Professor with Xiangtan University. His research interests include deep learning, intelligent information processing, and multi-sensor information fusion.

ZHIQIANG LI is currently pursuing the master's degree in control engineering with Xiangtan University. His research interests include image processing, pattern recognition, electric power fittings defect detection, and deep learning.

DONGBO ZHANG received the M.S. degree in computer application technology and the Ph.D. degree in control science and engineering from Hunan University, China. He has been a Professor with the College of Information Engineering, Xiangtan University, since 2006. His research interests include pattern recognition, image processing, machine learning, and machine intelligence.

LIANWEI TENG is currently pursuing the master's degree in control engineering with Xiangtan University. His research interests include image processing, pattern recognition, and deep learning.