applied sciences

MDPI

Article

Deep Learning-Based Bird's Nest Detection on Transmission Lines Using UAV Imagery

Jin Li, Daifu Yan, Kuan Luan, Zeyu Li and Hong Liang *

College of Automation, Harbin Engineering University, 150001 Harbin, China; lijin@hrbeu.edu.cn (J.L.); 619555782@hrbeu.edu.cn (D.Y.); luankuan@hrbeu.edu.cn (K.L.); zylee1@hrbeu.edu.cn (Z.L.)

* Correspondence: lh@hrbeu.edu.cn

Received: 5 August 2020; Accepted: 1 September 2020; Published: 4 September 2020

Abstract: In order to ensure the safety of transmission lines, the use of unmanned aerial vehicle (UAV) images for automatic object detection has important application prospects, such as the detection of birds' nests. The traditional bird's nest detection methods mainly include the study of morphological characteristics of the bird's nest. These methods have poor applicability and low accuracy. In this work, we propose a deep learning-based birds' nests automatic detection framework—region of interest (ROI) mining faster region-based convolutional neural networks (RCNN). First, the prior dimensions of anchors are obtained by using k-means clustering to improve the accuracy of coordinate boxes generation. Second, in order to balance the number of foreground and background samples in the training process, the focal loss function is introduced in the region proposal network (RPN) classification stage. Finally, the ROI mining module is added to solve the class imbalance problem in the classification stage, combined with the characteristics of difficult-to-classify bird's nest samples in the UAV images. After parameter optimization and experimental verification, the deep learning-based bird's nest automatic detection framework proposed in this work achieves high detection accuracy. In addition, the mean average precision (mAP) and formula 1 (F1) score of the proposed method are higher than the original faster RCNN and cascade RCNN. Our comparative analysis verifies the effectiveness of the proposed method.

Keywords: transmission line; bird's nest detection; convolutional neural network; deep learning

1. Introduction

With the increasing number of high-voltage transmission lines, damage caused by birds to the power systems is increasing. Some birds build their nests on the transmission towers. During rain, the birds' nests act as conductors and trigger power line tripping. Similarly, in a dry environment, the branches of the bird's nest are prone to fire, which not only affects the normal power supply, but also poses a huge security risk. Thus, in order to ensure the safe and reliable operation of power grid systems, and reduce the adverse effects of bird activities on transmission lines and other equipment, the research problem of detecting and locating bird nests on transmission lines and poles is of great scientific importance and practical significance.

Bird's nest detection on a high-voltage transmission line is a problem of image classification [1] and target-detection technology [2]. In literature, various methods for bird's nest detection on high-voltage transmission lines are presented. Most of these methods only consider the texture or color information of the bird's nest. This makes it impossible to locate the bird's nest accurately when the image's contrast is poor or the texture information is not rich.

In [3], authors propose a method that uses texture, color, shape and nest's area to find if the image contains a bird's nest. However, the bird's nest does not possess distinct characteristics. For instance, objects such as branches and grass also have strong texture characteristics, and have

the same color. This forms the basis of interference during detection. Authors in [4] present a method that extracts histogram of oriented gradient (HOG) features of birds' nests in images and used support vector machine for classification. However, the proposed method has poor adaptability and is not suitable for bird nest detection in texture rich environments. In [5], authors use the color and texture features for detecting the birds' nests on high-voltage transmission lines. The proposed method first identifies the tower, then uses the color feature of the bird's nest to determine the area of interest. Finally, the algorithms eliminate the interference area using the gray-level co-occurrence matrix feature. However, some of the inspection images of high-voltage transmission lines have poor contrast. This makes it difficult to identify the bird's nest on the basis of color and analyze texture characteristics. Therefore, it is not possible to use color and texture characteristics to describe the samples. Similarly in [6], authors propose a bird's nest detection method for high-speed railway catenary system. The authors use the histogram feature of the direction of the burr and the length of the burr to characterize the nest structure. Authors then use support vector machine (SVM) to identify and classify the birds' nests. The proposed method has a high detection rate, however, the application scenario of bird's nest detection in this paper is high-voltage transmission lines. It is noticeable that the images usually include insulators, towers and other interferences. In addition, the thickness of transmission wires in the tower area is uneven and the direction is variable. Therefore, this method is not applicable to the image samples used in this paper. In [7], the authors use a fully convolutional network with a novel pyramid structure to generate face proposals efficiently. In addition, online and offline hard sample mining are combined to further enhance the ability of networks. However, the joint use of hard sample mining is not restricted. It has great randomness, which is not conducive to the specific optimization of the model. In [8], the authors propose a new adaptive hard sample mining algorithm for person re-identification task. Through comprehensive comparison of the hard level differences between training batches and the differences in demand for hard sample numbers, the model is optimized and the accuracy is improved. However, due to the lack of pertinence, this method is not effective in detecting small-sized objects.

In recent years, breakthroughs have been made in the field of machine learning [9]. In the field of machine learning, deep learning has caused an unprecedented and tremendous impact [10]. Deep learning uses neural networks to solve linear inseparable problems. Deep learning uses large amounts of data and learns automatically without manual intervention.

Deep learning is widely used in target recognition [11] and multi-target detection [12]. Multiple feature maps are generated in the neural networks, each of which corresponds to many neurons. The features are extracted using a convolution filter [13]. The convolution operation [14] not only enhances the original features of signals but also reduce noise embedded in the imagery. For filters with the same step, the larger the image, the greater the number of neurons and the larger the number of weight parameters that need to be trained. Consequently, the training speed is low. Thus, the sample size is adjusted several times during the training process in order to improve the training efficiency and reduce the training overhead.

We use existing algorithms for object detection, and the accuracy is low. The main problem is that, during the model's training process, thousands of candidate regions may be generated from an input image. However, only a small number of these candidates contain the object of interest. This leads to a serious imbalance of the proportion between positive and negative samples, thus leading to class imbalance problem. Class imbalance leads the overall learning towards the useless, easily divided counterexample samples. Thus, resulting in invalid learning, i.e., only the background without objects can be distinguished, but the specific objects cannot be distinguished. Moreover, if the number of negative samples that are easy to classify are too large as compared to positive samples, it will have a negative impact on detection model optimization. In response to this problem, we propose a new automatic detection framework—region of interest (ROI) mining faster region-based convolutional neural networks (RCNN).

The main contributions of this work are as follows:

(1) We propose a new hard sample mining algorithm for object detection task. Through the comprehensive analysis of the datasets, the size characteristics of the detection objects are determined, and the limited conditions are given. The ROI mining method can focus better on the difficult-to-classify small-scale objects and optimize the model in a targeted manner.
(2) According to the characteristics of the annotation boxes of the datasets, we obtain adaptive prior anchors by using k-means clustering. This method reduces the convergence time of the model and improves the accuracy of the coordinate boxes generated.
(3) We combine the focal loss function in the model to solve the problem of class imbalance during the training process. This can improve the detection accuracy of the model.

The results presented in this work show that the proposed method achieves better results as compared to other methods proposed in the literature.

2. Faster Region-Based Convolutional Neural Networks (RCNN) in an Automatic Detection Framework

Region-based convolutional neural networks (RCNN) use a selective search to generate region proposals [15]. These region proposals are then used by CNN to extract features. Finally, these features are used by a support vector machine (SVM) to classify features in RCNN. The detection accuracy of the RCNN method on the pattern analysis, statistical modeling and computational learning visual object class (PASCAL VOC) dataset is much higher than the traditional methods [16], however, its training time and space overhead are huge. Each training image results in a large number of regions of interest (ROI). Moreover, we need to extract features for each ROI, and write the output to disk. Similarly, during testing, it is also necessary to extract ROI from the test samples and complete the detection after extracting the features from each ROI.

Faster RCNN is an improved algorithm based on RCNN and fast RCNN [17]. The main improvements of faster RCNN are: the SVM does not need to be trained; it combines multiple loss functions for a single level training process and for reducing time overhead; layers update during training and there is no need to write features to save disk space. The region proposal network (RPN) is proposed to generate candidate regions in a computationally efficient manner. It is noticeable that through alternate training, RPN and fast RCNN network share parameters which greatly improves the detection speed.

Faster RCNN consists of three parts: (1) feature extraction network, (2) region proposal network, (3) fast RCNN.

2.1. Feature Extraction Network

Faster RCNN is based on residual network (ResNet) [18] and consists of a series of residual blocks for extracting image features. The structure of the residual block is presented in Figure 1. Each residual block constitutes two paths: F(x) and x. The F(x) path fits the residual, and the x path presents an identity map. The addition requires that the dimensions of F(x) and x involved in the operation are same. ResNet effectively solves the problem of network degradation when the network depth increases. This means that as, the network's depth increases, the accuracy of the training set gradually decreases and the network performance deteriorates. ResNet mitigates this problem and improves the accuracy of object detection.

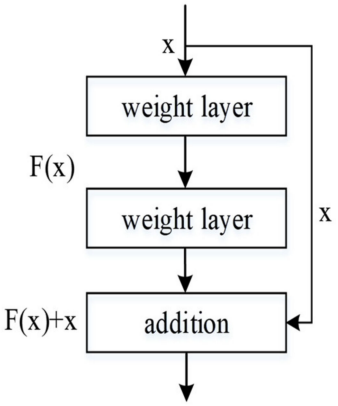

Figure 1. Residual block structure.

2.2. Region Proposal Network

RPN is a fully convolutional network capable of end-to-end training. The idea is to use a convolutional neural network to generate a set of rectangular object proposals, each with an objectness score. This is accomplished by the sliding window method on the feature map of the last shared convolutional layer to generate region proposals.

The small sliding window that is mapped back to the corresponding low-dimensional feature takes 3×3 window of the convolutional feature map as input. The features are inserted into next two 1×1 convolutional layers, namely, the regression layer, and the classification layer. Please note that the classification layer is used only for softmax classification and the regression layer is used for accurately locating the candidate regions. This is elaborated in, Figure 2.

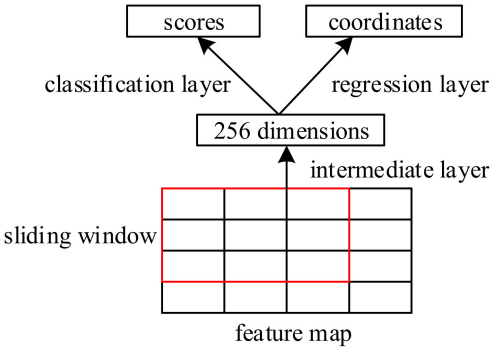

Figure 2. Region proposal network.

2.2.1. Anchor Mechanism

Anchor is the core of the RPN networks [19]. Anchors are boxes of preset size used to determine if any target objects are present in the corresponding receptive field at the center of each sliding window. Since the target size and the ratio of length to width are different, multiple scale windows are required. The anchor in faster RCNN sets the reference window size to 16, according to, (8, 16, 32) three multiples and three aspect ratios (1:1, 1:2, 2:1), i.e., a total of 9 scale anchors are obtained as presented in Figure 3.

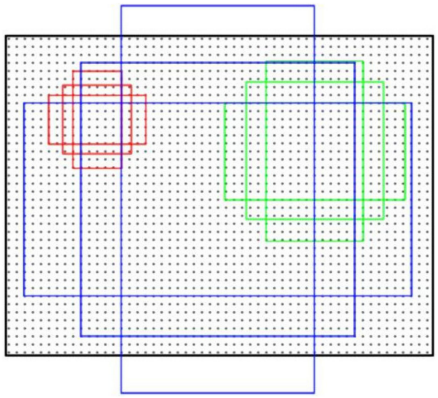

Figure 3. Multi-scale anchors.

2.2.2. Loss Function

The anchors with the largest intersection over union (IoU) with the ground truth boxes are considered as positive samples. Similarly, the anchors which have less than 0.3 IoU with ground truth boxes are considered as negative samples during RPN training. The loss function is defined as:

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}}\sum_i L_{cls}\left(p_i, p_i^*\right) + \lambda\frac{1}{N_{reg}}\sum_i p_i^* L_{reg}\left(t_i, t_i^*\right), \quad (1)$$

where, i and p_i represents the ith anchor in the mini-batch and the probability that the ith anchor is the foreground, respectively. When the ith anchor is the foreground, p_i^* is 1 and 0 otherwise. Please also note that t_i and t_i^* represents the coordinates of the predicted bounding box and ground truth coordinates, respectively.

2.3. Fast RCNN

During this phase, fast RCNN uses the obtained proposal feature maps, calculates the specific category of each proposal using the fully connected layer and softmax layer [20], and outputs the classification probability vector. In addition, the regression part uses bounding box regression to obtain the position offset for each proposal which is used to extract more accurate coordinates of the object's detection box.

3. Automatic Detection Framework-Related Work

3.1. Annotation Boxes Clustering

Faster RCNN uses nine default size anchor boxes for object detection. In this work, the statistical analysis of the size of annotation boxes reveals that the ground truth bounding box of the bird's nest dataset used in this paper are relatively small. In order to enhance the efficiency of bird's nest detection with different aspect ratios, we perform k-means clustering on bird's nest annotation boxes to find the sizes of the anchor boxes that are more suitable for the dataset. This is then used to design the anchors. Consequently, the convergence time of the model decreases, and the accuracy of bounding box coordinates is improved. When new images are added as input, the clustering method will automatically adjust the anchors according to the updated clustering result of the annotation boxes. This can adapt better to the characteristics of the dataset.

The k-means clustering method [21] comprises 3 steps. First initializing the number of categories and cluster centers. Second, calculating the distance between each bounding box and all cluster centers and assigning the bounding box to the nearest cluster. The mean value of every cluster is updated after each assignment. Finally, repeating the two aforementioned steps until the mean position of each cluster becomes stagnant.

In this work, we use 9 clusters for k-means clustering. The width and height of the ground truth bounding boxes are used as input features. The size distribution of bounding boxes in the bird's nest dataset after k-means clustering is presented in Figure 4. After clustering, the length and width of nine anchors are obtained. These are used as the basis for designing the anchors of the model in this paper.

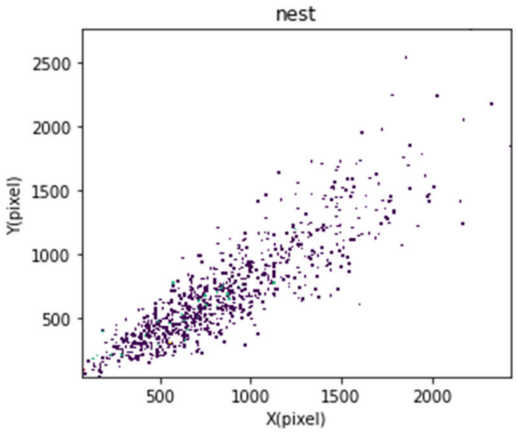

Figure 4. Annotation boxes clustering analysis. Please note that the abscissa X represents the width and the ordinate Y represents the height.

3.2. Focal Loss Function

In order to solve class imbalance [22] problem, this paper introduces focal loss function [23] in the RPN foreground and background classification stage, replacing L_{cls} in the RPN loss function.

The focal loss function is modified based on the standard cross entropy loss function that reduces the weight of samples that are easy to classify. Thus, the model is more focused on the samples that are difficult to classify during training.

In this work, focal loss is defined as:

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t), \tag{2}$$

where, α_t represents the weight ($\alpha_t \in [0,1]$) that is used to balance the uneven proportion of the number of positive and negative samples. γ is the focusing parameter ($\gamma \geq 0$), and $(1-p_t)^{\gamma}$ is called the modulation coefficient. The influence of the modulation coefficient on the loss value increases with the increase in γ. This solves the problem of imbalance between simple and complex training examples by reducing the weight of easily classifiable samples. Thus, the model is more focused on difficult and misclassified samples. p_t is defined as:

$$p_t = \begin{cases} p & y = 1 \\ 1-p & y = -1, \end{cases} \tag{3}$$

where, p represents the probability that the predicted sample belongs to class 1 ($p \in [0,1]$), and y represents the category of the label ($y \in \{\pm 1\}$). When the sample is misclassified, the value of p_t is very small, and $(1-p_t)$ tends to 1, so the loss of the sample is minutely affected. However, when the sample classification is correct, the value of p_t is large, and $(1-p_t)$ approaches to 0. Thus, the resulting sample loss value is very small. This loss function not only adjusts the weight of positive and negative samples, but also controls the weight of difficult and easy to classify samples. This solves the class imbalance problem in the RPN network during the training process.

4. Bird's Nest Detection Network Framework

Based on faster RCNN, this paper improves the network structure, and designs a network model based on ROI hard negative mining, called ROI mining faster RCNN. The proposed method effectively solves the problem of bird's nest detection for small objects with an imbalanced class problem.

4.1. Region of Interest (ROI) Mining Method

In training phase of the two-stage detection network, the number of negative samples may reach tens or even hundreds of times that of the positive samples. Most negative sample features correspond to the background in the receptive field of the input image. This is easy to classify, i.e., the classification loss value is small. Similarly, the positive set also contains many samples whose features are easy to classify. Therefore, during the training process, the decrease in the value of the classification loss function may be due to the correct classification of a large number of easy-to-classify samples. This means that the easy-to-classify samples dominate the decline in the value of the loss function, resulting in the final detector not effectively identifying the small bird's nest in complex scenes where the shapes are not obvious.

In response to the aforementioned problems, we propose the ROI mining method. Figure 5 presents the flow chart of ROI mining. The proposed method improves the proportion of small objects and difficult-to-classify objects in the classification loss and regression loss by screening and integrating all the ROI regions obtained through the RPN network. This method increases the proportion of small objects and difficult-to-classify objects in the total samples from 1:120 to 3:120. Thereby, a classifier is obtained that is more suitable for the efficient detection of difficult-to-classify nest samples with insignificant features.

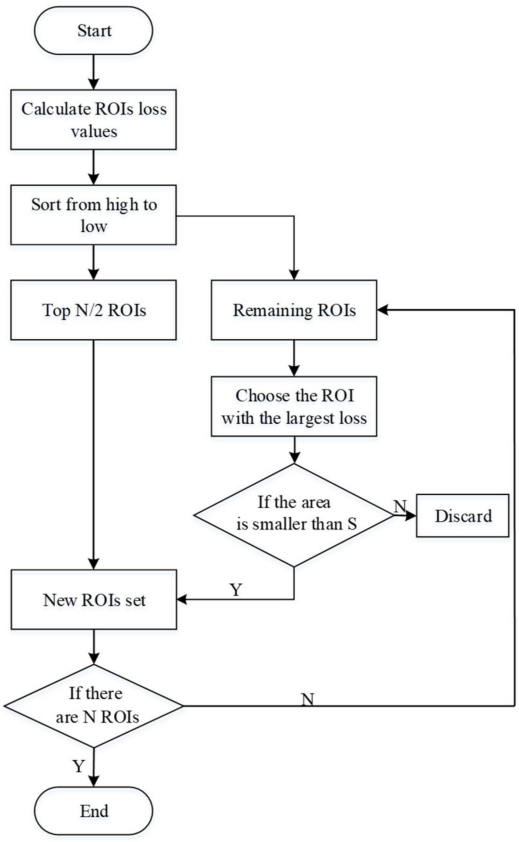

Figure 5. Region of interest (ROI) mining flowchart.

The detailed process of ROI Mining method is as follows:

Step 1. Before the training process starts, calculate the area of all the labeled boxes in the bird's nest dataset and sort the annotation boxes descending order. Afterwards, compute the median S using the areas.

Step 2. During the training process, obtain all the candidate ROIs that are to be used as an input to Fast RCNN classification loss and regression loss calculation. Calculate the corresponding total loss value for each ROI, and sort them according to the loss value from high to low.

Step 3. Select N numbers of ROIs to calculate loss. Place N/2 ROIs with the total loss values in the top of the set, and the remaining N/2 ROIs, whose loss value is high and area is smaller than S are selected.

Step 4. Use the replaced N ROIs as input for fast RCNN classification loss and regression loss calculation.

For the aforementioned process, we select N = 128 and apply this method to the bird's nest detection in this paper. This process makes small objects and difficult-to-classify samples dominate the loss function. Therefore, the class imbalance in the training process of the detector is mitigated, thereby improving the detection accuracy.

4.2. ROI Mining Faster RCNN Structure

In this work, we propose a method based on the 2-stage faster RCNN framework. We accomplish this by improving faster RCNN and proposing ROI mining faster RCNN. The ROI mining faster RCNN is based on ResNet-101 and is used for extracting features from images. This is done in order to ensure the accuracy of object detection in real-time applications and considering the detection problem in image samples obtained using an unmanned aerial vehicle (UAV). In order to improve the accuracy of bounding box coordinates generated by the algorithm, we use k-means clustering for extracting anchor boxes. In addition, we propose the focal loss function in RPN stage to balance the number of foreground and background samples. Moreover, we present the ROI mining module for solving the class imbalance problem during the training process. The overall flow chart is shown in Figure 6.

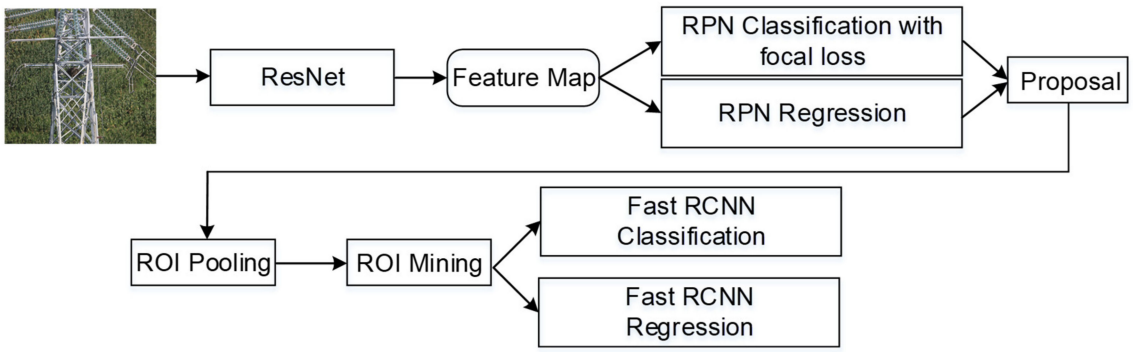

Figure 6. ROI mining faster region-based convolutional neural networks (RCNN) flow chart.

We use the annotated images as the input of the model, ResNet automatically extracts the features of the images and generates feature maps. The RPN network generates multiple candidate ROIs based on the information of the feature maps. Then a classifier is used to distinguish these ROIs into foreground and background, and a regression is used to make preliminary adjustments to the position of these ROIs. The processed ROIs are combined with the image information to obtain proposals, and input them into ROI pooling to obtain the output results of the same dimension. After that, we use ROI mining to select these results, input the selected special results into the final regression and classification networks. The method automatically generates the detection boxes, and finally obtains the detection maps.

The main content of this work is to solve the problem of class imbalance in the model training process. It is noticeable that the main cause of this problem is the error in the classification stage. In training, it is difficult for the classifier to recognize small-sized objects and it is easy to classify them as background, which leads to the imbalance between foreground and background. In the RPN stage, pixel-wise object detection, i.e., using the regression layer to locate, may cause errors. Most of these errors are caused by incorrect classification. Therefore, after solving the problems in the classification stage, we no longer consider the localization problems.

5. Simulation Results and Analysis

In this section, we present the results obtained using the proposed ROI mining faster RCNN detection framework and analyze the verification results.

5.1. Data Preparation

In this work, we collected 800 aerial images of electric towers. These images are acquired from Zhaotong power supply bureau of Yunnan province. We use these images for annotating bird's nest data for training and testing purposes. We use 5-fold cross-validation to evaluate the stability and detection performance of the models. In the 5-fold cross-validation experiment, the original dataset is randomly divided into five non-coincident sub-datasets, and then the models are trained and validated five times. Each time, four sub-datasets are selected as the training sets and one sub-dataset as the validation set. In the five training and validation processes, the sub-dataset used to validate the model is different each time. Finally, the average value of the five results is selected as the index representing the performance of the model. In addition, the training dataset images are subject to horizontal flip, vertical flip, and random rotation [24]. The images are randomly stretched within a certain range. Gaussian blur is applied, and salt-and-pepper noise are added to simulate the real-world environment. The uniform resource location (URL) of the dataset is shown in Supplementary Materials.

After the application of aforementioned preprocessing, we obtain 3000 images for training bird's nest. Figure 7 presents a schematic diagram of annotated pictures of a training set after data enhancement. Similarly, Figure 7a is an original image, and the rest of the images are partially enlarged from the original image using data processing.

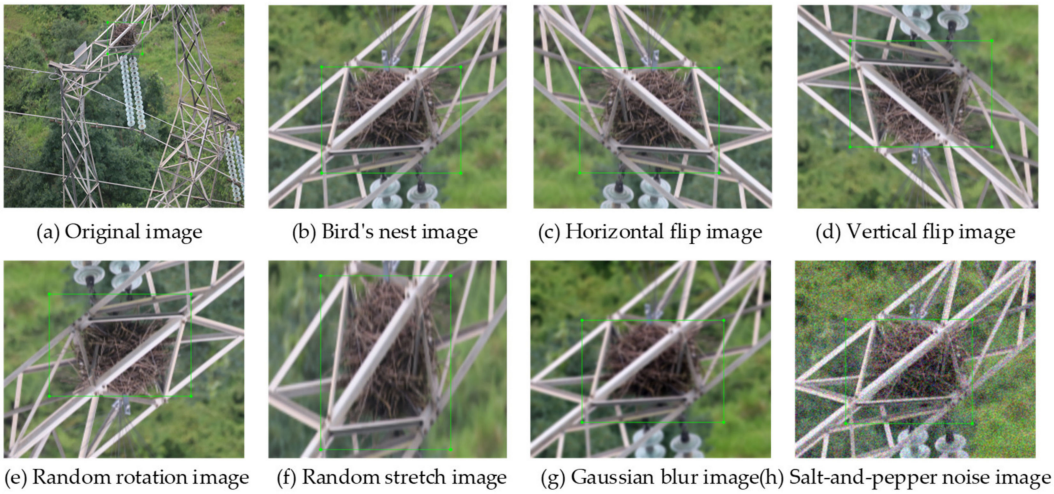

Figure 7. Schematic diagram of bird's nest samples after data enhancement.

5.2. Evaluation Index

In this work, we evaluate the faster RCNN, faster RCNN with focal loss, cascade RCNN and ROI mining faster RCNN models in terms of detection accuracy.

We use four evaluation parameters to evaluate the proposed work namely, precision, recall, F1 score and mean average precision (mAP). In the object detection task, we use the ratio of the area of the intersection between the final detection boxes and the sample annotation boxes to the area of their union to represent whether the final detection results are correct. The detected objects are divided into two categories, i.e., positive and negative. There are four cases for each category, i.e., true positive (TP), which correspond to correctly categorized samples in positive samples, i.e., the detector correctly detects the birds' nests as birds' nests; false positive (FP), samples that are incorrectly categorized into positive samples, i.e., the detector incorrectly detects the backgrounds as birds' nests; false negative (FN), negative examples that are incorrectly categorized into negative examples, i.e., the detector incorrectly detects the birds' nests as backgrounds; true negative (TN), negative examples correctly categorized, i.e., the detector correctly detects the backgrounds as backgrounds.

Based on this information, we define precision as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

Recall is defined as:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

F1 score is defined as:

$$\text{F1} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{6}$$

Since the detection object in this paper is only the bird's nest category, the AP value is the mAP value. It is expressed as:

$$\text{AP} = \int_0^1 \text{p}(\text{r})\text{dr} \tag{7}$$

where, p and r present the precision and recall, respectively.

5.3. Comparative Experiment

In this work, four models are used to train the bird nest recognition network on the dataset to verify the validity and efficiency of the proposed ROI mining faster RCNN model.

5.3.1. Simulation Setup

The feature extraction network used in this work is ResNet-101. In addition, we define the weight $\alpha_t = 1$, the focusing parameter $\gamma = 2$, the initial learning rate $\lambda = 0.001$, the training epochs is set to 20, and the batch size equals 1. Table 1 presents the basic configuration of the local computer. This configuration is independent of the detection accuracy in the experiment.

Table 1. Computer specifications.

Computer Configuration	Specific Parameters
CPU	Intel Core i7-8700k
GPU	NVIDIA GeForce GTX 1080Ti
Operating system	Ubuntu16.04, Canonical, London, United Kingdom
Random Access Memory	16 GB

5.3.2. Performance Evaluation

We present the comparison of the ROI mining faster RCNN model with faster RCNN, faster RCNN with focal loss, and cascade RCNN in terms of bird's nest detection using UAV images.

TP, TN, FP and FN are determined by the pixel-wise intersection-over-union. This work sets the intersection-over-union (IoU) threshold to 0.5. When the IoU of the output detection box and the

annotation box is greater than 0.5, we believe that the model correctly detects the object (P). Otherwise, it is considered that the model performs a wrong detection (F). The significance of TP, FP, FN and TN is to reflect the accuracy of detector classification and location, which directly affects recall and precision. Similarly, the accuracy of location is intuitively reflected by AP. In the detection, the real situations of TP, TN, FP and FN are shown in Figure 8.

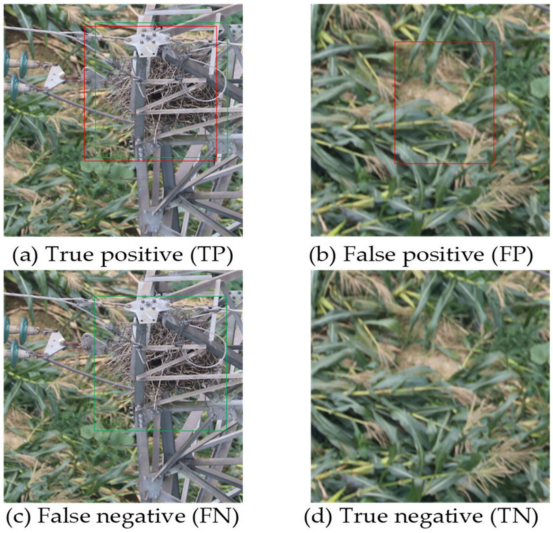

Figure 8. Practical examples of (**a**) true positive (TP), (**b**) false positive (FP), (**c**) false negative (FN) and (**d**) true negative (TN). Please note that the green box represents the label box and the red box represents the detection box.

In practical applications, in order to ensure the safe operation of transmission lines, it is necessary to accurately detect all birds' nests on the tower. Recall is used to represent the ability of the detector to detect birds' nests and suspected bird's nest objects. Excessive false detections result in the waste of computing resources, precision represents the accuracy of the detector to detect the birds' nests. F1 score is synthesized by precision and recall, which is used to balance the detection ability of the bird's nest and computing resources. AP is used to indicate the detection accuracy for bird's nest category. It measures whether the category and position of the bounding box predicted by the model are accurate. According to the analyses above, this work uses these four indicators to measure the detection ability of the detectors for the birds' nests.

This work uses 5-fold cross-validation to test the ROI mining faster RCNN. We evaluate the performance and stability of the model according to the mean and standard deviation of the validation mAP in the five validation sets. The results of the five tests are shown in Table 2.

Table 2. Five-fold cross-validation results.

K-Fold	Mean Average Precision (mAP)	Loss Value
1-fold	0.8249	0.2037
2-fold	0.8311	0.1756
3-fold	0.8243	0.2023
4-fold	0.8198	0.2103
5-fold	0.8256	0.1989
Mean	0.8251	0.1982
Standard deviation	0.0041	0.0133

As shown in Table 2, the mean mAP of ROI mining faster RCNN is 82.51%. After 5-fold cross-validation, the standard deviation of mAP is 0.0041. It can be seen from the validation standard

deviation and validation loss that ROI mining faster RCNN has performed better generalization ability and robustness.

In order to contrast and validate the functionality of our method in industrial applications, we use the images of high-voltage electric poles in the urban area as an input to detect foreign objects that are also small objects, e.g., honeycombs and water bottles. By detecting objects in different backgrounds, this can measure the effectiveness and stability of this method in a wide range of industrial applications. The results of the comparative experiments are shown in Table 3. It can be seen from the results that our method achieves high accuracy under different backgrounds and has good generalization.

Table 3. Comparison in different backgrounds.

Background	mAP
Foreign Object Detection	0.8011
Bird's Nest Detection	0.8251

In practical applications, some researchers uses the built-in functions of the software to automatically generate dataset labels. This greatly reduces the time for data preparation, however it affects detection accuracy of the model. We use the ground truth labeler function in MATLAB to label the training sets automatically, and compare the training result with the training result of the model using the manually labeled datasets. In this way, the recognition accuracy of the model under the automatically labeled datasets is verified. The comparison results are shown in Table 4.

Table 4. Comparison of labeling methods.

Labeling Method	mAP
Automatic labeling	0.6637
Manual labeling	0.8251

We use the enhanced dataset explained in Section 5.1 as the training set, and train the aforementioned networks. Table 5 presents the performance evaluation of the networks after testing is performed using the same test set.

Table 5. Detector performance.

Methods	mAP/%	F1/%	Recall/%	Precision/%
Faster RCNN	78.29	48.43	34.52	75.54
Faster RCNN with Focal Loss	78.99	48.87	35.72	75.76
Cascade RCNN	79.19	53.04	41.20	79.19
ROI Mining Faster RCNN	82.51	54.59	43.75	76.63

As presented in the results, the proposed method is able to achieve high detection accuracy for bird's nest detection. In addition, the proposed model ROI mining faster RCNN is able to outperform faster RCNN and cascade RCNN in terms of mAP, F1 score and recall. However, precision is slightly lower than cascade RCNN. It is noticeable that the proposed method of ROI mining faster RCNN improves recall rate of the object detection on the basis of classifying complex samples. Therefore, the detection accuracy enhances, while maintaining the precision. In terms of mAP, which reflects the overall performance of the detection network, the proposed method is able to cope with the inherent issues of the original model on the basis of focal loss that is added to faster RCNN. Similarly, the process of balancing the number of foregrounds and backgrounds during the training process improves mAP. The proposed method is able to achieve the mAP score of 82.51%, that proves the effectiveness of the ROI mining method.

It is evident from the results presented in this work that accuracy and mAP of ROI mining faster RCNN outperforms the faster RCNN and cascade RCNN methods. Moreover, precision of the proposed method is also comparable. Thus, the detection method proposed in this work greatly improves the process of bird's nest recognition on transmission lines and transmission poles. In addition, the proposed method successfully copes with the class imbalance problem during training and achieves the purpose of automatically and accurately detecting birds' nests in aerial transmission line images.

6. Conclusions

In this work, we consider aerial transmission line images as a research object, combined with the knowledge of image processing and neural networks. The work is based on the faster RCNN detection network, which is a deep learning-based detection method. In this paper, we propose ROI mining faster RCNN. First of all, we use the k-means clustering method to generate anchors coordinate boxes for bird's nest data. In addition, we deploy data enhancement strategies on the training set for improving the generalization and robustness of the proposed model. We propose focal loss function in the RPN stage to balance the proportion of positive and negative samples in the loss value, thereby alleviating the imbalance between the foreground and the background. Finally, we introduce the ROI mining method. The proposed model focuses on the classification of small object difficult-to-classify samples. It is noticeable that the loss value is dominated by the difficult-to-classify samples to solve the problem of class imbalance in the training process. We perform detailed comparative experiments and show that the proposed method improves the recall rate of object detection. Please note that while precision is slightly lower than cascade RCNN, at the same time, the mAP on the verification set reaches 82.51%, which is 4.22% higher than the original faster RCNN. The proposed method is able to accurately detect most of the bird's nest objects on aerial transmission lines. In addition, our results reveal that the proposed method is well adapted to the characteristics of aerial transmission line image datasets.

Supplementary Materials: The datasets are available online at https://zenodo.org/record/4015912#.X1O_0osRVPY.

Author Contributions: Conceptualization, J.L. and D.Y.; methodology, D.Y.; software, D.Y.; validation, D.Y.; formal analysis, D.Y.; investigation, D.Y.; resources, K.L. and D.Y.; data curation, D.Y. and Z.L.; writing—original draft preparation, D.Y.; writing—review and editing, D.Y.; visualization, D.Y.; supervision, J.L. and D.Y.; project administration, J.L. and H.L.; funding acquisition, J.L. and D.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (61773134, 61803117) and the Natural Science Foundation of Heilongjiang Province of China (YQ2019F003) and the Fundamental Research Funds for the Central Universities (3072020CFT0404).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Xuan, Q.; Chen, Z.; Liu, Y.; Huang, H.; Bao, G.; Zhang, D. Multiview Generative Adversarial Network and Its Application in Pearl Classification. *IEEE Trans. Ind. Electron.* **2019**, *66*, 8244–8252. [CrossRef]
2. Deng, Z.; Sun, H.; Zhou, S.; Zhao, J.; Lei, L.; Zou, H. Multi-scale object detection in remote sensing imagery with convolutional neural networks. *ISPRS J. Photogramm. Remote. Sens.* **2018**, *145*, 3–22. [CrossRef]
3. Xu, J.; Han, J.; Tong, Z.G.; Wang, Y.X. Method for detecting bird's nest on tower based on UAV image. *Comput. Eng. Appl.* **2017**, *53*, 231–235.
4. Duan, W.W.; Tang, P.; Jin, W.D.; Wei, P. Bird nest detection of railway contact net based on HOG feature in key areas. *China Railw.* **2015**, 73–77.
5. Jabid, T.; Uddin, M.Z. Rotation invariant power line insulator detection using local directional pattern and support vector machine. In Proceedings of the 2016 International Conference on Innovations in Science, Engineering and Technology (ICISET), Dhaka, Bangladesh, 28–29 October 2016; pp. 1–4. [CrossRef]
6. Wu, X.; Yuan, P.; Peng, Q.; Ngo, C.-W.; He, J.-Y. Detection of bird nests in overhead catenary system images for high-speed rail. *Pattern Recognit.* **2016**, *51*, 242–254. [CrossRef]
7. Zeng, D.; Zhao, F.; Ge, S.; Shen, W. Fast cascade face detection with pyramid network. *Pattern Recognit. Lett.* **2019**, *119*, 180–186. [CrossRef]

8. Chen, K.; Chen, Y.; Han, C.; Sang, N.; Gao, C. Hard sample mining makes person re-identification more efficient and accurate. *Neurocomputing* **2020**, *382*, 259–267. [CrossRef]
9. Stetco, A.; Dinmohammadi, F.; Zhao, X.; Robu, V.; Flynn, D.; Barnes, M.; Keane, J.; Nenadic, G. Machine learning methods for wind turbine condition monitoring: A review. *Renew. Energy* **2019**, *133*, 620–635. [CrossRef]
10. Voulodimos, A.; Doulamis, N.; Doulamis, A.; Protopapadakis, E. Deep Learning for Computer Vision: A Brief Review. *Comput. Intell. Neurosci.* **2018**, *2018*, 1–13. [CrossRef] [PubMed]
11. Guo, L.; Lei, Y.; Xing, S.; Yan, T.; Li, N. Deep Convolutional Transfer Learning Network: A New Method for Intelligent Fault Diagnosis of Machines with Unlabeled Data. *IEEE Trans. Ind. Electron.* **2018**, *66*, 7316–7325. [CrossRef]
12. Li, Z.; Dong, M.; Wen, S.; Hu, X.; Zhou, P.; Zeng, Z. CLU-CNNs: Object detection for medical images. *Neurocomputing* **2019**, *350*, 53–59. [CrossRef]
13. Zou, Q.; Zhang, Z.; Li, Q.; Qi, X.; Wang, Q.; Wang, S. DeepCrack: Learning Hierarchical Convolutional Features for Crack Detection. *IEEE Trans. Image Process.* **2019**, *28*, 1498–1512. [CrossRef] [PubMed]
14. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440. [CrossRef]
15. Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448. [CrossRef]
16. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Pdf ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
17. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
18. Wu, Z.; Shen, C.; Hengel, A.V.D. Wider or Deeper: Revisiting the ResNet Model for Visual Recognition. *Pattern Recognit.* **2019**, *90*, 119–133. [CrossRef]
19. Li, K.; Cheng, G.; Bu, S.; You, X. Rotation-Insensitive and Context-Augmented Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote. Sens.* **2017**, *56*, 2337–2348. [CrossRef]
20. Zhang, Y.-D.; Dong, Z.; Chen, X.; Jia, W.; Du, S.; Muhammad, K.; Wang, S.-H. Image based fruit category classification by 13-layer deep convolutional neural network and data augmentation. *Multimed. Tools Appl.* **2017**, *78*, 3613–3632. [CrossRef]
21. Hofmeyr, D. Degrees of freedom and model selection for k-means clustering. *Comput. Stat. Data Anal.* **2020**, *149*. [CrossRef]
22. Shrivastava, A.; Gupta, A.; Girshick, R. Training Region-Based Object Detectors with Online Hard Example Mining. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 761–769.
23. Shao, H.; Jiang, H.; Wang, F.; Zhao, H. An enhancement deep feature fusion method for rotating machinery fault diagnosis. *Knowl. Based Syst.* **2017**, *119*, 200–220. [CrossRef]
24. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [CrossRef] [PubMed]

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).