Real-time detection of UAV detection image of power line insulator bursting based on YOLOV3

Songbo Chen*, Chao Su b, Zhenxing Kuang c, Ye Ouyang*, Xiang Gong d

Guangdong Power Grid Corporation
*a songbo-chen@163.com, 860763-3468123, b 83456151@qq.com, 8613926683452
*b 125777382@qq.com, 860763-3468001, c 215904462@qq.com, 8613828567239
$d 490786274@qq.com, 8613750167875

Abstract: The insulator fault of high voltage line is the main factor of transmission accidents, so positioning and detection of burst insulator of power line has become an important part of routine detection. Traditional UAV detection is still mainly to evaluate the image transmitted by UAV manually, which is not only time-consuming, but also not accurate. In this paper, UAV aerial images are used to make datasets, the problem of insulator class imbalance (normal insulator and burst insulator) in the model training process is solved by using data augmentation, an improved depth learning algorithm is proposed based on YOLOv3, which provides rich semantic information for prediction layer by adding feature mapping module. At the same time, the residual network is introduced into the feature extraction, which improves the detection accuracy of small objects, more effectively extracts the object features of burst insulator, carries out real-time object detection and positioning, and completes the daily detection of insulator state of power transmission lines. The experimental results show that the detection accuracy (mAP) of the improved YOLOv3 algorithm is 91.22%, and the detection speed is 28 frames/s. The recall rate and Intersection Over Union (IOU) are also improved.

1. Introduction

In high-voltage transmission line, as an important part of power plant of overhead transmission line, insulator is used to fix bus bar and live conductor and provide enough conductor spacing, which plays an important role in electrical insulation and line support. Due to the long-term exposure of insulator to the external environment, it is often eroded by natural external forces, which is easily damaged, chipped and aged. Once the insulator fails, it will seriously affect the reliable transmission of high-voltage lines and the safe operation of power system. In order to ensure the safe and reliable operation of the whole transmission line, it is necessary to check the insulator of the transmission line in time and effectively find and remove the fault.

With the rapid growth of high-voltage transmission lines, the application of unmanned aerial vehicle (UAV) patrol technology in transmission line patrol is becoming more and more mature. Compared with the traditional manual patrol technology, UAV patrol technology can penetrate into the high-pressure work area which is difficult for many people to enter, greatly reducing the workload, shortening the patrol cycle, reducing the rate of false detection and missed detection. At the same time, using UAV as detection tool to locate component fault has become a new trend of high-voltage line detection. UAV patrol system consists of UAV air image acquisition and ground monitoring center data analysis, as shown in Figure 1. According to the patrol plan, the UAV airborne camera collects the
insulator images in different environments. UAV data processing center collects and analyzes data, and sends it to ground monitoring center through 4G communication network or memory card.

In order to detect and recognize insulators in aerial images under complex background, contour extraction, color feature, texture feature and machine learning are often used.

![Image of UAV detection system for transmission line](image)

**Figure 1. Structure of UAV detection system for transmission line**

Huang et al.\(^1\) used threshold segmentation technology to extract the features of insulator on the saturated component, but cannot eliminate the interference of other objects near the insulator, especially in the complex background. Markus\(^2\) used circular structural features to detect insulators. Zhai et al.\(^3\) used skeleton extraction technology to detect insulator. Yao et al.\(^4\) used insulator string feature extraction technology to detect insulator. However, these methods need to extract features manually, which requires a lot of work and low recognition rate. At the same time, due to the loss of some information, it is easy to cause the insulator's wrong or missing detection. A large number of insulator images acquired by UAV in complex background are difficult to recognize accurately.

In the past decade, depth learning has made remarkable progress, especially in the field of object recognition and object classification of computer vision. Alex net\(^5\) led a new wave of depth learning. Among all kinds of depth learning techniques, convolutional neural network (CNN) is the most popular depth learning network, which has been proved to be a powerful tool for image processing. Wang Wanguo\(^6\) et al. took the lead in using the Faster R-CNN\(^7\) algorithm in the component detection of UAV power detection image. The detection effect is good, but the real-time performance is poor. Chen Jingwen et al.\(^8\) put forward an aerial insulator detection method based on u-net depth network, which is an effective insulator detection method, but in the complex field background, its detection accuracy cannot meet the requirements. Du Fenglin et al.\(^9\) used RANSAC design missing insulator detection algorithm and SSD algorithm\(^10\) to effectively locate the missing insulator in the image, but the robustness of fitting straight line to judge the defect is insufficient. In recent years, many object detection methods based on CNN have achieved great success, such as two-stage object detection algorithm using classification and regression, such as R-CNN\(^11\), Faster R-CNN\(^12\) and R-FCN\(^13\). However, in the process of pixel classification of insulator image by CNN method, each pixel needs to select an image classification block in advance. For insulator image with complex background, the image classification blocks of two adjacent pixels have high similarity and redundancy, which easily leads to slow network training speed. Recently, more and more attention has been paid to the exploration of object detection methods that are more suitable for the small depth neural network structure of embedded devices, such as the single-stage algorithm of direct regression for object detection, such as YOLO\(^14\), SSD\(^15\), DSSD\(^16\), RSSD\(^17\), FSSD\(^18\), YOLOv3\(^19\) and Retina Net\(^20\). However, the models of YOLO, YOLOv2\(^21\) and YOLOv3 are still very large, which are difficult to apply to most embedded devices with limited memory. In order to solve the above problem that the model is too large, Tiny YOLO\(^22\) reduces the size of the model by simplifying the network structure, which enhances the real-time performance of the algorithm with only a small loss of detection accuracy. At the same time, in order to achieve the best balance between the object detection accuracy and the real-time embedded requirements, in document\(^23\), the Fire micro structure and SSD in the end-to-end object recognition network structure SqueezeNet\(^24\) are combined to form Tiny SSD, so as to reduce the model size and improve the detection accuracy. In the above methods, the detection performance of SSD is relatively good, at the same time, its detection speed is fast, and its accuracy is also high. However, because SSD has a smaller visual range and a lower object detection accuracy, the detection performance of small objects is poor. In order to solve these problems, some researchers proposed DSSD and YOLOv3 in the top-down structure to improve the performance of small object detection.
Due to the low resolution of the object, the occlusion of the object, the illumination change of the insulator and other problems in the UAV scene, the amount of image feature extraction is small, which is the bottleneck to restrict the deployment of the depth learning model and SSD algorithm of the mobile terminal (or the object terminal).

This paper mainly studies a real-time method to detect insulator defects in the transmission line image taken by high-definition UAV in complex background. The collected insulator data set is labeled manually and accurately, and the position of damaged insulator is directly located by end-to-end lightweight depth neural network training. In this paper, the residual SqueezeNet network is properly reduced and improved. As the basic network of YOLOv3, this network includes Depth-wise Separable Convolution\cite{23} and ResNet\cite{24}. By extracting insulator features of different scales and adding feature mapping module to provide rich semantic information for the prediction layer, the detection accuracy of small objects is improved, the object features of burst insulator are effectively extracted and real-time object detection and positioning is carried out. The experiment shows that, while maintaining high accuracy, the storage capacity of the model is reduced, the deduction speed of high-definition insulator image is greatly improved, and the problem of real-time detection of insulator defects on transmission line by UAV is guaranteed to be solved.

2. Improved YOLOv3 model based on SqueezeNet

2.1 YOLOv3 model structure

YOLOv3 is the latest version after YOLOv1 and YOLOv2. Its basic idea is: first, the feature extraction network extracts features from the input image to obtain a feature image with specific size, such as 13*13, and then divides the input image into 13×13 grid cells; second, if the central coordinate of a object in the ground truth is in which grid cell, the grid cell will predict the object. Then, through RPN network, each grid cell will predict a certain number of Bounding Boxes. Of course, only the Bounding Boxes with the largest IOU of the ground truth are used to predict the object. The network structure consists of the skeleton network Darknet 53 and the detection network, which are used for feature extraction and multi-scale prediction respectively. Among them, Darket 53 is composed of convolution layer and residual layer. The feature extraction model combines the advantages of YOLOv2, Darknet-19 and ResNet, uses more convolution layers of 3*3 and 1*1, and adds some Shortcut Connection structures at the back. The output layer uses convolution layer instead of full connection layer, discards softmax classifier in classification prediction, and uses 53 convolution layers in total. In order to solve the imbalance of positive and negative sample proportion in the single structure direct regression method (such as SSD, YOLO, etc.) with preset fixed anchor boxes, YOLOv3 uses Focal loss to reduce the weight of negative samples by sacrificing the detection accuracy.

The multi-scale prediction of YOLOv3 will be carried out on the feature map with the size of 52*52, 26*26 and 13*13. But before the feature map outputs the prediction results, the feature fusion operation will be carried out first, and the features with high semantic low resolution and low semantic high resolution will be spliced together, so that the features with high resolution also contain rich semantic information. The specific fusion process is shown in Figure 2 below: first, perform five convolution operations on the 13*13 size feature graph, the convolution kernel size is 1*1, the step size is 1, and the number of convolution kernels is 1024. The convolution layer with convolution kernel size of 3*3, step size of 1 and convolution kernel number of 512 is reconnected to achieve the effect of dimensionality reduction. After the above-mentioned 13*13 feature map is cross convoluted, the feature map of 26*26 is formed by double UpSample, and then it is spliced(tensor connected) with its upper level feature (feature size 26*26) to complete local feature interaction. Repeat the above operations to complete the splicing fusion of 52*52 size feature map. Finally, the output size of the three YOLO interaction layers is 52*52, 26*26, 13*13 respectively. Then, the detection is carried out on the three different scale feature images for classification and location regression. This method can improve the detection effect of small objects significantly. YOLOv3 uses logistic regression to calculate whether each bounding box contains objects (two categories: Yes or no). At the same time, it
uses independent logistic classifier instead of the previous softmax to predict the class probability. This algorithm solves the shortcomings of traditional YOLOv1 and YOLOv2 algorithms, such as coarse particle detection and weak detection of small objects. Through positioning and category prediction on multi-scale feature map, the detection effect of small objects is improved significantly.

In the process of calculating the network loss, YOLOv3 divides all prediction boxes into positive samples (intersection over Union with the actual dimension box area > 0.5) and negative samples (intersection over Union with the actual dimension box area < 0.5). In general, the proportion of the object in the picture is far less than that of the background (especially the burst insulator on the high-voltage transmission line), so the number of positive and negative samples is quite different. Moreover, most of them are easily classified negative samples (belonging to background samples), which makes the training process unable to fully learn the information belonging to those classified samples, and may also cover up the role of other classified samples. At this time, the network loss function converges slowly in the iterative process of a large number of simple samples and is difficult to reach the optimal. To solve this problem of unbalanced number of difficult and easy classification samples, this paper improves the standard cross entropy loss function by using Focal-loss. The traditional cross entropy loss function is as follows:

$$CE(p, y) = \begin{cases} -\log(p) & \text{if } y = 1 \\ -\log(1-p) & \text{otherwise} \end{cases} \tag{1}$$

In the formula, $p$ is the probability of prediction, and its value is between 0-1; $y$ is the true label. It can be seen that the larger the output probability of positive samples is, the smaller the loss is, and the smaller the output probability of negative samples is, the smaller the loss is. This paper modifies two aspects: 1) control the effect weight of the class imbalance problem on the loss function by adding a weight factor; 2) change the contribution weight of difficult-separate samples and easy-separate samples to the loss function. The modified formula is:

$$FE(p, y) = \begin{cases} -\alpha (1-p') \log(p) & \text{if } y = 1 \\ -((1-\alpha) p') \log(1-p) & \text{otherwise} \end{cases} \tag{2}$$

In the formula, $\alpha$ can be set to 0.25 to control the category imbalance, and $\gamma$ is usually set to 2 to adjust the rate of reducing the weight of easy-separate classify samples, so that the model pays more attention to difficult-separate classify samples in training.

The object detection network loss function of YOLOv3 is shown in formula (3).

$$Loss\_Function = \lambda_{noobj} \sum_{i=0}^{s^2} \sum_{j=0}^{B} \left[ \gamma \left( y_{ij} - \hat{y}_{ij} \right) \right]$$

$$+ \lambda_{coord} \sum_{i=0}^{s^2} \sum_{j=0}^{B} \left[ \left( x_{ij} - \hat{x}_{ij} \right)^2 + \left( y_{ij} - \hat{y}_{ij} \right)^2 \right]$$

$$+ \lambda_{coord} \sum_{i=0}^{s^2} \sum_{j=0}^{B} \left[ \left( w_{ij} - \hat{w}_{ij} \right)^2 + \left( h_{ij} - \hat{h}_{ij} \right)^2 \right]$$

$$+ \lambda_{coord} \sum_{i=0}^{s^2} \sum_{j=0}^{B} \left[ \left( -C_{ij} \log \hat{C}_{ij} - (1-C_{ij}) \log(1-\hat{C}_{ij}) \right) \right]$$

$$+ \lambda_{coord} \sum_{i=0}^{s^2} \sum_{j=0}^{B} \left[ \left( -\alpha C_{ij} (1-\hat{C}_{ij}) \log \hat{C}_{ij} - (1-\alpha) (1-C_{ij}) \log(1-\hat{C}_{ij}) \right) \right]$$

$$+ \lambda_{coord} \sum_{i=0}^{s^2} \sum_{j=0}^{B} \left[ \left( -\alpha C_{ij} (1-\hat{C}_{ij}) \log \hat{C}_{ij} - (1-\alpha) (1-C_{ij}) \log(1-\hat{C}_{ij}) \right) \right]$$

In the formula, $s^2$ is all grid cells of output feature map, $B$ is the number of predicted frames of each grid, $i$ is the $i$-th grid, $j$ is the $j$-th box predicted by this grid, $obj$ is with objects, $noobj$ is no objects, $C_{ij}$ is the category confidence of real objects, $\hat{C}_{ij}$ is the category confidence of predicted objects, $p_{ij}(\hat{C}_{ij})$ is the probability of real box category, $p_{ij}(\hat{C}_{ij})$ is the probability of predicted box category, $\gamma$, $\alpha$ judges whether the j-th box in the i-th grid is responsible for the prediction of objects; $\lambda_{coord}$, $\lambda_{obj}$, $\lambda_{noobj}$ and $\lambda_{cls}$ is the penalty factor. The loss function consists of five parts. The first part indicates the loss caused by the central coordinate error of the prediction and the actual box; the second part indicates the loss caused by the width and height error of the prediction and the actual box; the third part indicates the loss of the confidence of the box, through calculating the IOU value.
(Intersection over Union) of the prediction box and the actual box; the fourth part is the loss caused by the box confidence of the corresponding object not detected; the fifth part is the loss of classification, and the classification loss is calculated when the grid contains the object.

2.2 SqueezeNet model

SqueezeNet is a lightweight and efficient CNN model proposed by Han et al. It has fewer parameters than AlexNet, and the accuracy of single model is close to AlexNet. The model is mainly compressed and optimized from the following four aspects: (1) model compression: the pre-trained model is compressed, and network pruning and quantification are adopted to make it into a small model. (2) CNN microstructure optimization: to optimize the design of a single convolution layer, a large number of 1x1 small convolution kernels are used to replace the 3x3 convolution kernels, reducing the parameter amount by 9 times. (3) CNN macrostructure optimization: optimization design at the network architecture level, network depth (layers), bypass connection, etc., such as reducing the number of input channels of the 3x3 convolution kernel, reducing the convolution kernel parameters, decomposing the convolution layer into the Squeeze layer and expand layer, and encapsulating it into the Fire Module. (4) design space optimization: different super parameters, network chain structure, optimizer and other combination optimization, such as the use of delay down sampling technology, adding feature layer module, making the convolution layer has a larger perception field of vision.

The basic unit of SqueezeNet network adopts modular convolution, commonly known as Fire Module. Fire Module mainly includes two layers of convolution operations, and each of them connects a ReLU activation layer: one is the squeeze layer with 1x1 convolution kernel, which contains all U 1*1 convolution kernels; the other is the expand layer with 1x1 and 3x3 convolution kernels, which contains E1 1*1 and E3 3*3 convolution kernels, and meets the requirements of U < E1 + E3. The SqueezeNet model has nine layers of Fire Module, three max pooling layers are interspersed in the middle, and the last layer uses average pooling instead of full connection layer, which greatly reduces the number of parameters.

The operation process of the fire module is shown in Figure 3. The size of the input feature map is $H \times W \times M$. First, the input feature map is compressed through the squeeze layer to obtain the output feature map with the size of $H \times W \times S_{in}$. At this time, the size of the feature map remains the same, but the overall size is reduced from M to S1x1. The output feature maps of the compression layer are dispersed into the expansion layer to form 1x1 convolution kernel and 3x3 convolution kernel respectively, and then the convolution results are fused and connected. Finally, the size of the feature map is unchanged, but the number of channels is changed to E1x1+E3x3.
However, SqueezeNet has two disadvantages: low classification accuracy and high computational complexity. Although the structural parameters are very small, it is not conducive to deployment on mobile devices.

In order to reduce the computational complexity of the structure, the paper [25] proposed the Wide Fire Module structure inspired by four lightweight computing models, including SqueezeNet [16], ReNext [15], MobileNet [20], ShuffleNet [21]. In WFM, they introduced group convolution to replace the 3 x 3 and 1 x 1 extended convolutions in the Fire Module. ShuffleNet [21] proposes a group convolution algorithm, which effectively reduces the parameters and computation of SqueezeNet structure. Its main idea is to improve ResNet by using group convolution and channel shuffle. In ShuffleNet, use the grouping method. In order to solve the problem that channel information of different groups cannot be shared, channel shuffling technology is introduced to shuffle channels of different groups.

In order to better reduce the computational complexity and improve the accuracy, this paper combines channel shuffle and WFM, and proposes a Fire Module method based on channel shuffle and mixed group convolution [27] to replace WFM, as shown in Figure 3. Mixed group convolution can flexibly design multiple super parameters through the following parameters to balance the number of features generated by mixed group convolution:

1) Group size \( g \): it determines how many different types of kernels are used to represent input channels. In the extreme case of \( g = 1 \), Mix-Grouped is equivalent to a normal depth convolution. Generally, the efficiency and accuracy of the model are further improved when \( g \) is chosen from 1-5.

2) Kernel size of each group \( K \): theoretically, each group can have any kernel size. However, if two groups have the same convolution kernel size, they can be combined into one group, so we can assume that each group has a different kernel size. At the same time, because the small kernel usually needs less parameters and triggers, we stipulate that the size of each group of kernels always starts from 1x1 and increases monotonously by 2. In other words, the kernel size of group \( l \) has always been \( 2^{l+1} \). For example, three sets of mixed group convolutions always use kernel size \{1x1, 3X3, 5x5\}. Under this constraint, the kernel size of each group can be pre-determined by the group size \( g \), thus simplifies the design process.

3) Channel size of each group \( C \): the general method of channel grouping includes: (1) equal grouping: each group has the same number of filters; (2) exponential grouping: group \( l \) accounts for about \( 2^l \) of the total number of channels. For example, given that the total filter size of four groups of mixconv is 32, the channel is divided into \( (8, 8, 8, 8) \) equal partition and \( (16, 8, 4, 4) \) exponential
2.3 Improved YOLOv3 model structure

Although the feature extraction ability of convolutional neural network is improving with the deepening of network layers, model size and model prediction speed still need to be considered in practical engineering. Depth neural network contains dozens or even hundreds of layers of network, which has a large number of weight parameters. Saving and calling weight parameters has a high demand on the performance of equipment. The traditional YOLOv3 adopts the self defined backbone network Darknet-53, which has complex model calculation, high storage requirements and low real-time performance, and is not conducive to the detection of mobile terminals.

In order to solve the above problems, this paper proposes a lightweight neural network model for real-time object detection (as shown in Figure 5). On the basis of the traditional YOLOv3 network, the mixed group Fire Module (MGFM) shown in Figure 4 will be used to replace the Fire Module in SqueezeNet (as shown in Figure 3), and the improved SqueezeNet will be associated with YOLOv3 to design the YOLOv3-Squeeze network, which can improve the inference speed of image recognition.

The improved mixed group Fire Module (MGFM) in SqueezeNet network is used as the modular convolution (as shown in Figure 4) to replace the residual module in the darknet-53 backbone network. Inspired by MobileNet, the standard convolution of the Fire Module layer 2 is replaced by the depth-wise separable convolution. Channel shuffle and mixed group convolution are used to solve the problem of "poor information flow" caused by depth-wise separable convolution, as shown in Figure 4. The depth-wise separable convolution realizes the separation of channels and regions, breaks the interaction between the number of channels and the size of convolution kernel, mainly including the Depthwise convolution and the Pointwise convolution. The Depthwise convolution is responsible for filtering, and the Pointwise convolution (also 1x1 convolution) is responsible for channel conversion.

The depth-wise separable convolution can only extract features from the corresponding feature map. Therefore, each channel feature map can be reused by pointwise mixed group convolution. Compared with the standard convolution, the compression ratio of the number of parameters is:

\[
\frac{D_x \cdot D_y \cdot M \cdot D_z}{D_x \cdot D_y \cdot M \cdot N} = \frac{1}{N} + \frac{1}{GD_x}
\]  

(4)
Among them, $D_k$ represents the size of convolution kernel, $D_F$ represents the feature size of input, $M$ represents the number of input channels, $N$ represents the number of output channels, and $G$ represents the size of groups in pointwise group convolution. The denominator part represents the parameter of standard convolution, and the numerator part represents the parameter of depth-wise separable convolution.

![Figure 5. Improved squeezet framework using fixed group Fire Module](image)

3. Burst insulator testing experiment

This experiment was carried out on a computer with Intel Core i7-9750H (2.60GHz) CPU, NavidiaGeForce RTX2070GPU and 16GB RAM under Ubuntu 18.04LTS. In this test, 4856 training samples of various background insulators from China Southern Power Grid were carried out, including 4256 images of damaged insulators and 600 images of normal insulators. In this paper, online data enhancement is used to expand the data set. Before the data set is sent to the network for training, the images will be enhanced immediately, including illumination change, rotation, flipping, cutting, translation, etc., which makes the image scene of the input network have a variety. According to the ratio of 4:1, the training set consists of 3885 images, and the test set consists of 971 images.

For the training set and test set of this experiment, the graphical image annotation tool "LabelImg" is used for manual annotation, which is determined by the coordinates of the upper left corner and the lower right corner. The results of these labels are converted to TFRecord file format, a simple record oriented binary format commonly used in Tensorflow and Keras. The insulator box is a rectangular area covering the entire insulator area.

3.1 Evaluation index

3.1.1 Evaluation of mean average precision (mAP)

The model evaluation of object detection problem often needs to be carried out from two aspects of classification and positioning and both of them are dispensable. Conventional model evaluation indexes include accuracy (the ratio of correctly predicted positive samples to total positive samples), precision (the ratio of correctly predicted positive samples to all predicted positive samples) and recall (the ratio of correctly predicted positive samples to total real positive samples). Set $R_{cd}$ to represent the correct sample detection rate, $N_c$ to represent the number of defects correctly identifying insulator bursting, $N_o$ to represent the number of insulator bursting incorrectly identifying, $I_o$ to represent the intersection of prediction frame and real marking frame of bursting insulator, and $U_o$ to represent the union between prediction frame and real marking frame.

$$R_{cd} = \left(1 - \frac{N_c}{N_c + N_o} \right) \times 100\%$$

$$mAP = \frac{I_o}{U_o}$$

Different from the traditional description of mAP50, this paper uses mean average precision (mAP) to evaluate the detection accuracy of burst insulator, and adopts the method of pointwise interpolation.
for average to represent the mAP according to different thresholds. First, the average recognition accuracy $\bar{\text{P}}_{\text{IOU}}$ is obtained by counting the Intersection over Union $P_{\text{IOU}}$ of each detection frame.

$$\bar{\text{P}}_{\text{IOU}, r} = \frac{\sum_{i=1}^{N} P_{\text{IOU}, r}}{N}$$

(6)

In the formula: $N$ is the number of detection frame.

According to the different threshold $r$ values of recall rate, the maximum precision value corresponding to $\text{mAP}_r$ is calculated respectively, and then the final measurement index $\text{mAP}$ can be obtained by summing. Among them, $r$ corresponds to 11 levels of 0%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% and 100% respectively.

$$\text{mAP} = \frac{1}{11} \sum_{r} \text{P}_{\text{IOU}, r}$$

(7)

3.1.2 Intersection Over Union (IOU)

IOU (Intersection Over Union) is a concept used in object detection. It is a standard to measure the accuracy of detecting the corresponding object in a specific data set. It is also a very important function as object detection algorithm performance mAP calculation. The function can be represented as follows:

$$\text{IOU} = \frac{\text{Area of overlap for Graph A and B}}{\text{Area of Union for Graph A and B}} = \frac{A \cap B}{A \cup B}$$

(8)

Among them, A and B represent the original label box (the rectangular box surrounding the object), B represents the candidate box (the rectangular box surrounding the object) predicted by the algorithm, and A and B intersect with each other. In short, IOU is the result of dividing the overlapped part of two regions by the set part of two regions.

3.1.3 Precision-Recall Curve

In PR curve, Recall is the X-axis, which reflects the learning model's ability to cover positive cases, while Precision is the Y-axis, which reflects the accuracy of learning model's prediction of positive cases. According to the prediction results of model learning, the accuracy rate and recall rate of each position are calculated. The curve reflects the trade-off between the accuracy of positive case recognition and the ability of positive case coverage. In training evaluation, the relationship between PR can be adjusted by setting threshold, which is also called classification boundary value. When score > threshold, the classification is positive, and when score < threshold, the classification is negative; when threshold increases, accuracy rate increases, recall rate decreases; when threshold decreases, accuracy rate decreases, recall rate increases; accuracy rate and recall rate are two variables that are mutually contained and contradictory, which cannot be increased at the same time.

3.1.4 F1-Measure

F1-Measure, also known as F-score, is a common evaluation standard in the field of information retrieval. The calculation formula is:

$$F_{\beta} = \frac{(\beta^2 + 1)PR}{\beta^2P + R}$$

(9)

Where $P$ stands for precision, R stands for recall, and $\beta$ stands for different evaluation criteria. For example, $\beta=1$ stands for the most common F1-Measure. At this time, R and P are equally important (with the same weight). When $\beta=2$, F2-Measure values recall more, whereas when $\beta=0.5$, precision has a higher weight than recall. $\beta$ can take any non-negative real number. As shown in the figure, as the number of samples increases, the F value gradually converges.
3.1.5 Speed Evaluation

Another important performance index of the object detection algorithm is speed. Only when it is fast, it can be applied in real-time, which is of great significance for the realization of UAV online detection in the future. The commonly used measurement of detection speed is Frame Per Second (FPS), that is, the number of images that can be processed per second. The equation is:

$$FPS = \frac{1.0}{\text{TotalTime} / \text{FrameCount}}$$

(10)

3.2 Test results

Comparing the algorithm proposed in this paper with the four mainstream frameworks in the field of object detection (Faster R-CNN, SSD, R-FCN and YOLOv3), it can be seen from table 1 that Faster R-CNN uses RPN network structure to extract a large number of insulator candidate frames that may be damaged, which makes the accuracy of insulator defect detection high and the recall rate high, but the process of RPN extracting candidate frames takes a lot of time, resulting in its slow detection speed and cannot be used in the real-time detection process. The accuracy of yolov3 is 93.75%, but other performance indexes are weak (such as recall rate and average recognition accuracy). When SSD and R-FCN maintain high accuracy, their recall rate and average recognition accuracy are better than YOLOv3. The algorithm proposed in this paper can achieve the best defect detection, and achieve a good balance between accuracy and recall rate. At the same time, it can improve the frame processing speed and meet the real-time processing needs.

| mainstream framework | algorithm | Insulator Broken Defect |
|-----------------------|----------|-------------------------|
|                       |          | Precision               |
| Faster R-CNN          | Resnet-50| 89.5                    |
|                       | Resnet-101| 89.6                   |
| SSD                   | Resnet-50| 14.7                    |
|                       | Resnet-101| 44.9                   |
| YOLOv3                | Darknet-53| 90.4                   |
| R-FCN                 | Resnet-101| 93.68                  |
| The Proposed method   | darknet-53| 91.03                  |

For the detection of burst insulator, when the threshold value of IOU is set as 0.5, the value of mAP@0.5 is 91.22%. Figure 7 shows the test results of some insulators in the test set.

The proposed method can achieve a good result of 91.03% accuracy and 83.56% recall rate. The corresponding PR curve is shown in Figure 8 below. It can be seen from the figure that when the recall
rate changes from 0 to 85%, the average accuracy mAP is still above 60%, which shows that the method achieves a good balance between the selection of average accuracy and recall rate.

Figure 7. Test results of some insulators in the test set

Figure 8. PR curve of the algorithm proposed in this paper

4. Conclusion
Aiming at the requirement of real-time detection of transmission line insulator damage with high-definition UAV aerial image, this paper proposes a defect detection method based on the latest YOLOv3 improvement. Under the premise of maintaining the insulator monitoring performance, the network model of this method adopts the lightweight SqueezeNet network to provide rich semantic information for the prediction layer by adding the feature mapping module. At the same time, the residual network is introduced into the feature extraction to improve the detection accuracy of small objects. The real-time detection of insulator state of transmission line has an average accuracy of 91.22%, and the detection speed is more than 28 frames / second.

REFERENCES
[1] X. Huang and Z. Zhang, “A method to extract insulator image from aerial image of helicopter patrol,” Power System Technology, no. 1, pp. 194–197, 2010.
[2] M. Oberweger, A. Wendel, H. Bischof, Visual, recognition and fault detection for power line insulators, in: Proc. 19th Computer Vision Winter Workshop, 2014, pp. 1–8.

[3] Zhai Yongjie, Wang Di, Wu Yang, Cheng Haiyan, “Two-stage Recognition Method of Aerial Insulator Images Based on Skeleton Extraction”, Journal of North China Electric Power University, Vol.42, No.3, p:105-110. 2015.

[4] Yao Chun-yu, Jin Li-jun, Yan Shu-jia, “Recognition of Insulator String in Power Grid Patrol Images”, Journal of System Simulation, Vol 24, No.9. p: 1818-1822, 2012.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” Communications of the Acm, vol. 60, no. 2, p. 2012, 2012.

[6] Wang Wanguo, Tian Bing, Liu Yue. Research on recognition of small power components in UAV patrol image based on RCNN [J]. Journal of Geo-Information Science, 2017 (2)

[7] Wei Liu, Dragomir Angelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single shot multibox detector. In European conference on computer vision, pages 21–37. Springer, 2016.

[8] Chen Jingwen, Zhou Xin, Zhang Rong. Aerial insulator detection based on u-net network [J]. Journal of Shaanxi University of Science and Technology, 2018 (4)

[9] Du Fenglin. Application of object detection algorithm based on depth learning [D]. Anhui University of Science and Technology, 2018

[10] R. Girshick, “Fast R-CNN,” in IEEE International Conference on Computer Vision (ICCV), 2015.

[11] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: towards real-time object detection with region proposal networks,” in International Conference on Neural Information Processing systems, 2015, pp. 91–99.

[12] Jifeng Dai, YiLi, Kaiming He, Jian Sun. “R-FCN: Object Detection via Region-based Fully Convolutional Networks” In CVPR, 2016

[13] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 779–788, 2016.

[14] ChengYang Fu, Wei Liu, Ananth Ranga, Ambrish Tyagi, Alexander C Berg, “DSSD : Deconvolutional Single Shot Detector”, arXiv:1701.06659v1, [cs.CV] 23, jan 2017.

[15] Jisoo Jeong, Hyojin Park, Nojun Kwak. “Enhancement of SSD by concatenating feature maps for object detection”, arXiv:1705.09587, [cs.CV],2017

[16] Zuoxin Li, Fuqiang Zhou, “FSSD: Feature Fusion Single Shot Multibox Detector”, arXiv:1712.09960v1, [cs.CV], 17, May 2018

[17] Joseph Redmon, Ali Farhadi. “YOLOv3: An Incremental Improvement”, arXiv:1804.02767, [cs.CV],sun, 8,2018

[18] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, Piotr Dollár, “Focal Loss for Dense Object Detection”, arXiv:1708.02002, [cs.CV], 2018.

[19] Joseph Redmon, Ali Farhadi. “YOLO9000: Better, Faster, Stronger”. arXiv:1612.08242v1 [cs.CV] 25 Dec 2016

[20] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi. “You Only Look Once: Unified, Real-Time Object Detection”. arXiv:1506.02640v5 [cs.CV] 9 May 2016

[21] Alexander Wong, Mohammad Javad Shafiee, Francis Li, Brendan Chwyl. “Tiny SSD: A Tiny Single-shot Detection Deep Convolutional Neural Network for Real-time Embedded Object Detection”. arXiv:1802.06488v1 [cs.CV] 19 Feb 2018

[22] Iandola, Forrest N., et al. “SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 1MB model size.” arXiv preprint arXiv:1602.07360 (2016).

[23] Chollet F. Xception: deep learning with depthwise separable convolutions[C]/Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017:1800-1807.

[24] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[C]/Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.
[25] Hengfui Liau, Nimmagadda Yamini, YengLiong Wong, “Fire SSD: Wide Fire Modules based Single Shot Detector on Edge Device”, arXiv:1806.05363, [cs.CV], 11 Dec, 2018
[26] Ting Zhang, Guo-Jun Qi, Bin Xiao, Jingdong Wang, “Interleaved Group Convolutions for Deep Neural Networks”, arXiv:1707.02725, [cs.CV], 18, Jul, 2017.