Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105 (Online)
Volume **8**, Number **2**, May **2023**

Bird's Nest Detection Method of Transmission Tower based on Improved YOLOX-S

Ren-Jie Song

Department of Computer Science
Northeast Electric Power University
No. 169 Changchun Road, Jilin, Jilin, China
1939811347@qq.com

Li-Peng Xu*

Department of Computer Science
Northeast Electric Power University
No. 169 Changchun Road, Jilin, Jilin, China
Corresponding Author: 2117876117@qq.com

Received July 15, 2022, revised September 31, 2022, accepted November 9, 2022.

Abstract. *Birds nesting on power transmission towers is prone to trigger power outage accidents such as grounding of foreign objects. At present, aerial photos of transmission towers by UAV are often used for bird's nest detection. However, the complex background of field aerial photography and the small size of the bird's nest target make detection more difficult. The existing mainstream target detection algorithm cannot accurately detect the bird's nest at a high detection speed. To solve this problem, an improved YOLOX-S bird's nest detection method for transmission towers is proposed in this paper. YOLOX-S is a target detection algorithm based on deep learning. In the process of improvement, the accuracy of bird's nest detection is improved by adding the CBAM attention mechanism, increasing the receptive field of the backbone network characteristic map, and introducing the EIOU loss function. Several groups of comparative experiments were constructed to prove the effectiveness of the method. Experimental results show that the mAP of the improved YOLOX-S algorithm improves 4.6% compared with the original algorithm, and can effectively detect the bird's nest of the transmission tower in real-time.*
Keywords: bird's nest detection, YOLOX-S, deep learning, attention mechanism, loss function

1. **Introduction.** For the electric transmission system, the one of the main dangerous elements is bird's nest [1]. Birds dropping nesting materials in transmission towers can cause power outages due to line tripping, and excrement can contaminate insulators and cause flashover. Therefore, regular cleaning of bird's nest on transmission towers is critical to the safe operation of transmission lines. Transmission towers have large targets [2] and often located in remote mountainous areas. Bird nests are generally located at the top of the tower where it is difficult to observe. The traditional manual inspection method has high intensity, low efficiency and the effect cannot be guaranteed [3]. With the development of UAV technology and deep learning algorithm, it is possible to detect bird nests on transmission towers by remote controlled UAV. The research of bird's nest detection algorithm with good real-time performance and high identification accuracy is important for scientific control of bird damage and maintenance of the stable operation of the power grid.

Deep learning algorithms used in bird nest detection can be divided into two categories: namely two-stage and one-stage methods [4]. The two-stage detection algorithm is mainly represented by Faster R-CNN [5], R-FCN [6], Mask R-CNN [7], Cascade R-CNN [8]. Such methods will first generate candidate regions that may contain detection targets (Region Proposal) [9] when detecting the input image. Candidate regions were classified and location regression was performed to obtain the detection results. The one-stage detection algorithm is mainly represented by SSD [10], RetinaNet [11], YOLO [12], YOLOV3 [13]. Such methods do not need to generate candidate areas that may contain bird nests. Get the bird's nest category and position coordinates directly. The one-stage detection algorithm is optimized on the basis of the two-stage detection algorithm. With its advantages of simple structure and efficient computation, it has attracted the attention of many researchers [14] and developed rapidly.

Because the scene of aerial photography is relatively complex, it is easy to interfere with various factors. And the regions having bird nests are relatively small and not easy to be noticed [15]. How to accurately and efficiently detect bird nests on transmission towers is an urgent problem to be solved. Li et al. [16] used Faster R-CNN algorithm to detect bird's nest and replaced the backbone feature network with ResNet-50. When extracting the initial candidate regions, the K-means method was used to cluster the candidate regions of the bird's nest, and three sliding windows of different sizes were used to adapt to objects of different sizes in the RPN. However, the steps were too complicated for real-time detection. Zhao et al. [17] used Cascade R-CNN algorithm to detect bird's nests and added a multi-stage structure to R-CNN. In the network, samples with corresponding quality are selected through three different IOU thresholds for training and detection. The IOU value of the candidate box is increased step by step so that the candidate box at the previous level can adapt to the next stage with a higher IOU threshold. Compared with Faster RCNN, the precision of Cascade R-CNN is greatly improved, but the speed is much slower. Zhong et al. [18] proposed to optimize the structure of the YOLOV3 network with the idea of DenseNet convolutional neural network. The loss function of the prediction category imbalance and the width and height loss function of the prediction box is redefined. Considering that the bird's nest is small and frequently appears, the maximum feature map is changed to 104×104. Nevertheless, training time is long, and the speed of bird's nest detection is still not ideal. Satheeswari et al. [19] proposed to use VGG16 and EfficientNetB7 to extract image features and generate feature maps, and verified that EfficientNetB7 had a high accuracy as the SSD feature extraction network. But it also increases the accuracy of detection by adding a large number of parameters.

YOLOX-S [20] is a one-stage target detection algorithm proposed by Megvii Technology in 2021. Excellent accuracy and speed in the COCO public dataset compared to other mainstream detection algorithms. Moreover, YOLOX-S adopts an anchor-free mechanism, which improves the detection ability of small targets. However, it is found through experiments that there is still a large space for improving the accuracy when detecting bird's nest on the transmission tower, and there is a certain deviation between positioning the target box and the real target box. To solve the above problems, this paper proposes a bird's nest detection method for transmission towers based on improved YOLOX-S. While maintaining rapid detection, the accuracy of bird's nest detection is greatly improved, which is more consistent with the real target box size. This is important for efficient and accurate detection of bird nests on transmission towers and smooth operation of transmission equipment.

2. **Bird's nest detection based on improved YOLOX-S.** Compared with original YOLOX-S algorithm, the improved YOLOX-S algorithm uses CBAM attentional mechanism to process low-level branches of feature fusion to focus on key features. SPP structure in the backbone network was modified to increase the receptive field of feature map and enhance feature extraction ability. Replace the IOU loss function to speed up network convergence and improve regression accuracy.

2.1. **YOLOX-S target detection algorithm.** YOLOX-S is optimized on the basis of YOLOv5-S, integrating many excellent strategies and methods. YOLOX-S includes four parts: input image, feature extraction (backbone), feature fusion (neck) and prediction of corresponding objects. Mosaic [21] and MixUp [22] data enhancement techniques were used to input images. Use Mosaic random zooming, random cropping, and random placement to splice images together. It greatly enriches the background of detecting objects. Mixup is an enhancement strategy based on Mosaic, in which two filled images are weighted and fused to create a new image.

Based on Darknet53, the backbone feature extraction network introduces CSP [23] module to prevent the generation of repeated gradient information. Focus structure is used to reduce information loss during down-sampling. SPP layer increased the receptive field by max-pooling of the feature layer [24], and the SiLU activation function is used as the activation function.

The feature fusion part receives features provided by the three-layer branches of the backbone network. And forms FPN + PAN multi-scale feature fusion structure by referring to FPN [25] (Feature Pyramid Networks) and PANet [26] networks. FPN is fused from the bottom up through up-sampling, and PAN is fused from the top down through down-sampling to improve the information on different scales. Activation function also adopts SiLU activation function.

In object prediction, large, medium, and small targets are predicted on feature maps of different proportions. The coupled head was replaced by the decoupled head. The decoupled head is divided into three parts, which are used to judge the type of the target, judge whether the target is the foreground or background, and predict the coordinate information of the target box respectively. Different branches are adopted according to the different contents concerned by classification and positioning, which improve the speed of network convergence. Anchor-free mechanism and SimOTA were used for optimization. The anchor-free mechanism can directly predict coordinate value and height and width value without clustering analysis of data. Eliminates extra computational and training overhead of the algorithm in the post-processing stage [27]. The use of SimOTA can intelligently analyze how many positive samples each ground truth should have and dynamically match positive samples for targets of different sizes. The network structure of YOLOX-S is shown in Figure 1.

2.2. **CBAM attentional mechanism.** Target detection sometimes neglects important feature information and reduces detection accuracy, which can be solved by introducing an attention mechanism in a convolutional neural network. CBAM (Convolutional Block Attention Module) is a lightweight attention mechanism module with two independent sub-modules. They are respectively CAM (Channel Attention Module) and SAM (Spatial Attention Module), which can assign weight from multiple dimensions, making up the shortcoming of these two methods that only focus on channel dimension or spatial dimension, and makes the output information focus on more critical feature areas. CAM structure is shown in Figure 2, and SAM structure is shown in Figure 3.

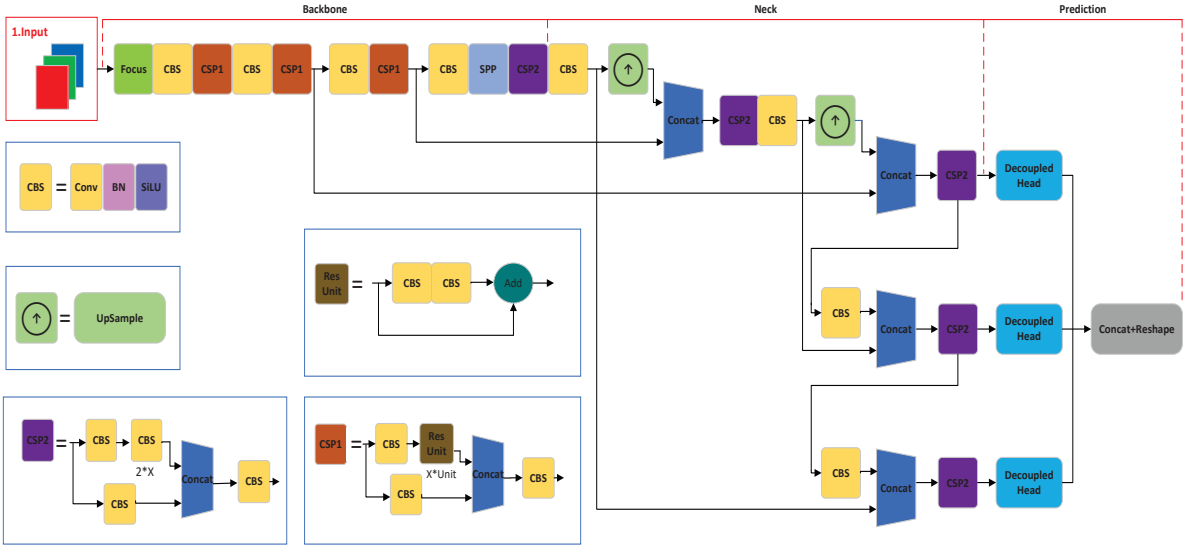


FIGURE 1. YOLOX-S network structure

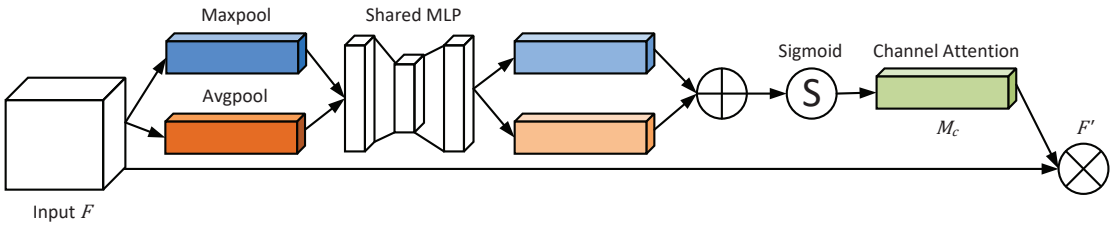


FIGURE 2. Channel attention module structure

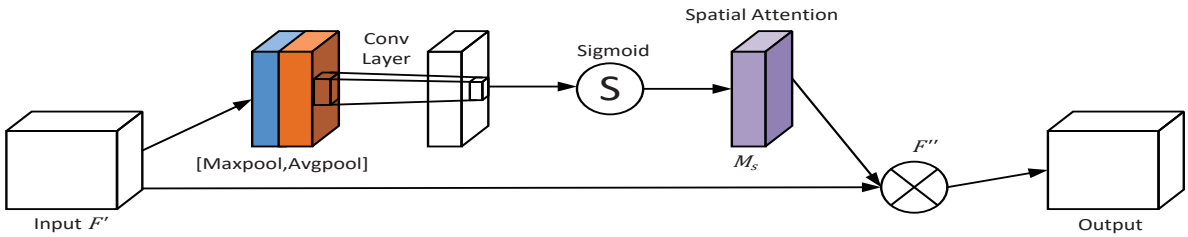


FIGURE 3. Spatial attention module structure

Input F gets output F' through channel attention module, and F' as input passes through the spatial attention module to get output F''.

$$F' = M_c(F) \otimes F \tag{1}$$

$$F'' = M_s(F') \otimes F' \tag{2}$$

Where $\otimes$ represents the multiplication of corresponding elements, the feature map $F \in R^{C\times H\times W}$, and the channel attention map $M_c \in R^{C\times 1\times 1}$

$$M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \tag{3}$$

$MLP()$ is a multi-layer perceptron with weight sharing. $MaxPool()$ and $AvgPool()$ represent max-pooling and average-pooling operations respectively. $\sigma()$ represents sigmoid activation function. In the channel attention module, the max-pooling and average-pooling operations based on height and width are respectively adopted for input feature map F. The result is added after two full join operations, and the weight of each channel was obtained by the sigmoid activation function. Finally, the feature map F' is obtained by multiplying it with the feature map F, and the subsequent generated spatial attention map $M_s \in R^{1\times H\times W}$is:

$$M_s(F') = \sigma(f^{7\times 7}([MaxPool(F'); AvgPool(F')])) \tag{4}$$

Where $f^{7\times 7}$represents the convolution operation of the 7×7 convolution kernel and [;] represents concatenate operation. In the spatial attention module, max-pooling and average-pooling operations based on height and width of input feature map $F'$are also adopted respectively. After stacking the two feature layers, 7×7 convolution kernel is used for dimensionality reduction, and the weight of each feature point is obtained through sigmoid activation function. Finally, the feature map F'' is obtained by multiplying the feature map F'.

2.3. **SPP structure.** SPP structure, also known as spatial pyramid pooling, was initially proposed to solve the problem that the input image size cannot be changed during training. In the past, stretching and clipping to a uniform image size resulted in distorted images and lost information, while SPP structure can generate feature vectors of fixed size for arbitrary input size and reduce the occurrence of over-fitting.

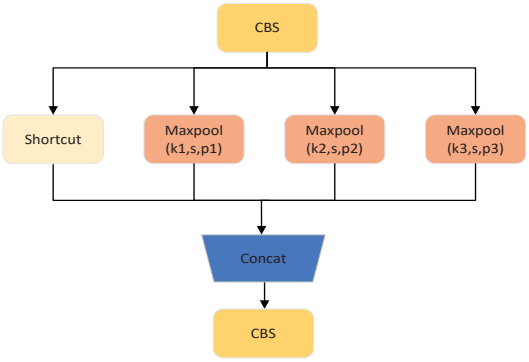


FIGURE 4. SPP structure

The SPP module in YOLOX-S draws on the spatial pyramid idea, which is used to increase the receptive field of feature map of backbone network, extract spatial feature information of different sizes, and improve the robustness of the model. It mainly consists of four parallel branches, including three max-pooling operations and a shortcut. Where k1, k2 and k3 are convolution kernels with sizes of 5×5, 9×9 and 13×13 respectively; s is step size with sizes of 1; p1, p2 and p3 are padding with sizes of 2, 4 and 6 respectively. First, the input channel was halved through convolution, and the feature maps of different receptive fields were generated by max-pooling with convolution kernels of different sizes, which maintain the original size. Then, different feature maps can fuse local and global features through concatenate operation, which enriches the expressive ability. After convolution again, the number of channels is adjusted to the original state, reducing the amount of calculation. The SPP structure is shown in Figure 4.

2.4. **EIOU loss function.** The loss function is used to evaluate the difference between the predicted value and the real value of the model, so the loss function is very important in training. Generally, the better the loss function is, the faster the convergence rate is, and the better it is to get an excellent model. IOU loss function is used in YOLOX-S to calculate the loss of the reg part. By taking the boundary box formed by 4 points as a whole, regression is carried out to calculate the ratio of intersection and union between the prediction box and real box to evaluate the effect. The formula of IOU and IOU loss function is as follows:

$$IOU = \frac{|A \cap B|}{|A \cup B|} \tag{5}$$

$$L_{IOU} = 1 - \frac{|A \cap B|}{|A \cup B|} \tag{6}$$

A is the prediction box, and B is the real box. EIOU loss function is improved on CIOU loss function. CIOU loss function solves the defects of the IOU loss function, but produces new defects.

$$L_{CIOU} = 1 - IOU + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha v \tag{7}$$

$$v = \frac{4}{\pi^2}(arctan\frac{w^{gt}}{h^{gt}} - arctan\frac{w}{h})^2 \tag{8}$$

$$\alpha = \frac{v}{(1 - IOU) + v} \tag{9}$$

Where, $\rho(\cdot)$ indicates the Euclidean distance, α is the trade-off parameter, v is used to measure the consistency of aspect ratio, b and b^{gt} represent the center point of the prediction box and real target box respectively, w^{gt} and h^{gt} represent the width and height of the target box respectively, w and h represent the width and height of prediction box respectively. The gradient of v with respect to w and h is:

$$\frac{\partial v}{\partial w} = \frac{8}{\pi^2}(arctan\frac{w^{gt}}{h^{gt}} - arctan\frac{w}{h}) \times \frac{h}{w^2 + h^2} \tag{10}$$

$$\frac{\partial v}{\partial h} = -\frac{8}{\pi^2}(arctan\frac{w^{gt}}{h^{gt}} - arctan\frac{w}{h}) \times \frac{w}{w^2 + h^2} \tag{11}$$

It can be seen from the definition of the CIOU loss function that its measurement of aspect ratio is too complex, resulting in slow convergence. And aspect ratio describes the relative value, there is a certain ambiguity. EIOU loss function replaces the αv part, including three parts: overlap loss, center distance loss and width and height loss.

$$L_{EIOU} = 1 - IOU + \frac{\rho^2(b, b^{gt})}{c^2} + \frac{\rho^2(w, w^{gt})}{C_w^2} + \frac{\rho^2(h, h^{gt})}{C_h^2} \tag{12}$$

Respectively c, C_w and C_h are the diagonal lengths, width and height of the smallest enclosing box covering the two boxes. EIOU loss function divides aspect ratio and calculates the length and width of the target box respectively, which makes up the defect of CIOU loss function.

2.5. **Improved YOLOX-S target detection algorithm.** Shallow network has strong ability to represent geometric details, while deep network has rich semantic information. Firstly, the Dark3 branch is processed by CBAM attention mechanism, and the channel weight is adjusted appropriately to focus the spatial geometric information and reduce the interference of background factors. Concatenate with the deep features after bottom-up fusion to improve the detection accuracy of the bird's nest. Referring to the idea of residual network [28] structure, shortcut is used to process the original Dark3 incoming features to prevent network degradation, and the next layer branch is introduced for feature fusion. Secondly, the SPP structure is improved in the YOLOX-S backbone feature extraction network. For the results after max-pooling of 5×5 convolution kernel and 9×9 convolution kernel, the smaller 5×5 convolution kernel is used for max-pooling operation. Without increasing too much computation, the receptive field of feature map can be further improved to obtain more global information. Multi-scale feature maps after max-pooling are fused to strengthen the feature extraction ability of the network and make the feature fusion part better use the features extracted from the backbone network. The improved SPP structure is called SPPH structure, as shown in Figure 5.

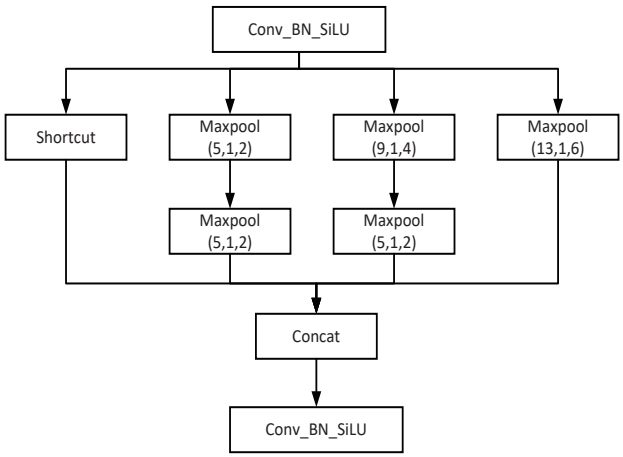


FIGURE 5. SPPH structure

On this basis, the loss function replaces the IOU with EIOU. IOU function has the following problems: When boxes A and B do not intersect, loss is always 0, which cannot reflect the distance between the two, and further learning and training cannot be achieved through gradient return. Moreover, the prediction box and the real box cannot accurately reflect the size of the overlap degree between the two boxes. EIOU solves the above problems and improves the convergence speed and regression accuracy of the mode. The network structure of improved YOLOX-S is shown in Figure 6.

3. Experiment.

3.1. **Data set processing.** In this paper, there were 612 bird's nest images of the transmission line inspected by UAV. In order to prevent over-fitting in training and increase the robustness of the model, the data set was expanded to 1360 images by adding noise and random flipping. According to the experimental requirements, 80% of them are used as training sets and 20% as test sets. During training, bird's nest pictures less sheltered by transmission towers are selected as far as possible. LabelImg software was used to manually generate labels for all images, and an XML label file with PascalVOC format was generated. The label file mainly contains the size of the picture, the category of the object, the coordinates of the upper left corner (x_{min}, y_{min}) and the lower right corner

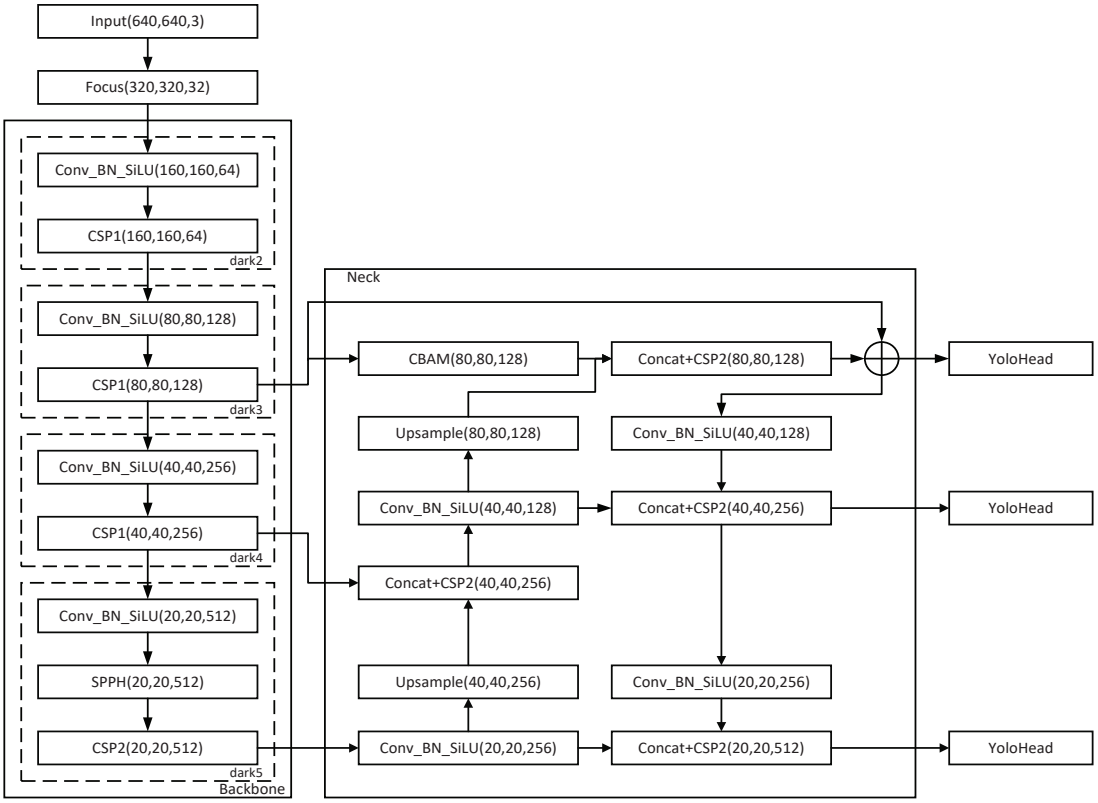


FIGURE 6. Improved YOLOX-S structure

(x_{max}, y_{max}) of the rectangular frame where the bird's nest is located in the picture. To facilitate training, the images were uniformly converted to a size of 1250 x 1250. The augmented example is shown in Figure 7.

FIGURE 7. Data augmentation

3.2. **Experimental environment.** Considering the deep learning framework pytorch concise features of efficient and good compatibility, all experiments are based on this framework. The personal computer CPU used in the experiment is Intel(R) Core(TM) i7-10750H, GPU is NVIDIA Geforce GTX 2060, video memory is 6G, cuda version is 10.0, cudnn version is 7.4.1, operating system is Windows 10, batch size is set to 4, confidence is set to 0.5, initial learning rate is 0.001, weight decay is 0.0005.

3.3. **Experimental results and comparative analysis.** In order to prove the effectiveness of the method, the following three groups of experiments were designed and

compared on the same training set, test set and hardware. In experiment 1, only CBAM attention mechanism was added, which was called YOLOX-SC algorithm. In Experiment 2, SPP structure was modified to SPPH structure on the basis of experiment 1, which was called YOLOX-SCH algorithm. In Experiment 3, IOU loss function was modified to EIOU loss function on the basis of experiment 2, which was called YOLOX-SCHE algorithm. YOLOX-SCHE algorithm is the final improved YOLOX-S algorithm. The loss comparison curve of the above algorithms and the original YOLOX-S algorithm is shown in Figure 8.

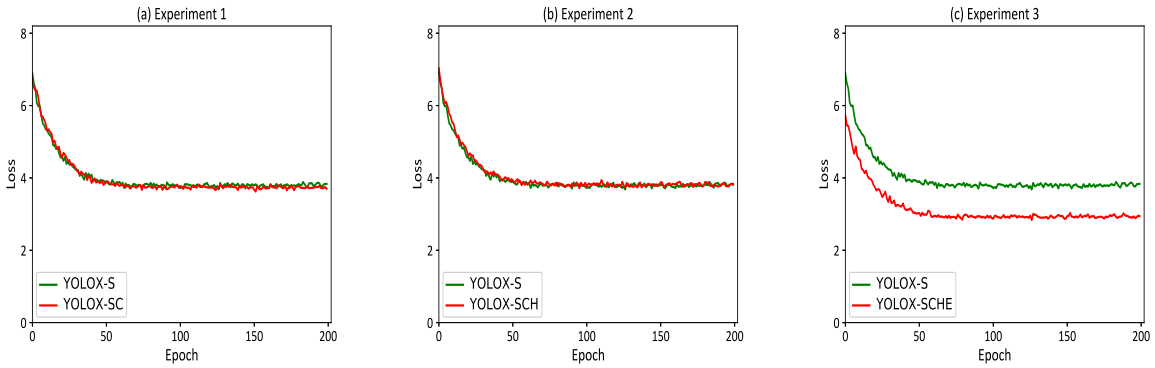


FIGURE 8. Comparison of loss in different experiments

In the experiment, mAP (Mean Average Precision) and FPS (Frames Per Second) are important evaluation indexes to measure model performance. Where mAP represents the average value of multiple AP (Average Precision) to measure the accuracy of the detection process. AP represents the area covered by the P-R (Precision-Recall) curve. Precision represents the percentage of all predicted targets that are correct, and Recall represents the percentage of actual targets that are correct.

$$Presion = \frac{TP}{TP + FP} \tag{13}$$

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

Among them, the number of positive samples correctly divided is called TP, the number of negative samples incorrectly divided into positive samples is called FP, and the number of positive samples incorrectly divided into negative samples is called FN. FPS can measure the detection speed of the model, and only fast detection speed can meet the real-time requirements.

TABLE 1. Comparison of different methods in mAP and FPS

Method	mAP(%)	FPS
YOLOV5-S	88.3	57.6
YOLOX-S	90.6	55.4
YOLOX-SC	91.1	54.1
YOLOX-SCH	94.5	52
YOLOX-SCHE(ours)	95.2	51.5

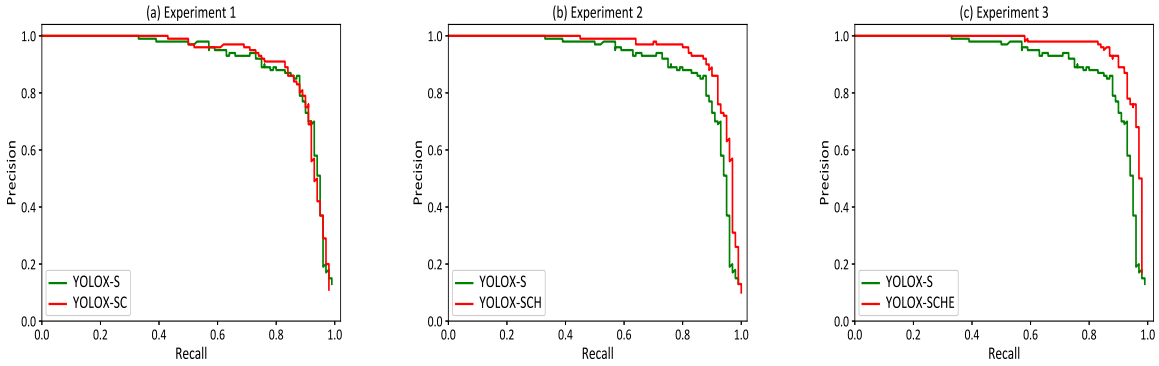


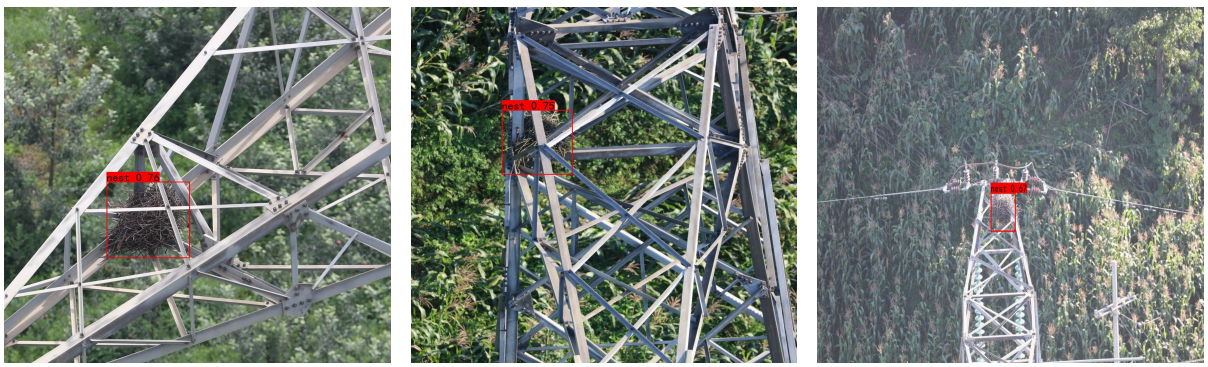
FIGURE 9. Comparison of P-R curves in different experiments

It can be seen from Table 1, Figure 8, and Figure 9 that the mAP of the algorithm shows a steady upward trend after gradual modification, loss convergence tended to be stable after 50 epochs, and FPS decreased slightly. In experiment 1, the accuracy of YOLOX-SC was improved by 0.5%, LOSS was decreased by 0.05 and FPS was decreased by 1.3 compared with YOLOX-S. In experiment 2, the accuracy of YOLOX-SCH was improved by 3.9%, LOSS increased by 0.04 and FPS decreased by 3.4 compared with YOLOX-S. In experiment 3, YOLOX-SCHE improved accuracy by 4.6%, LOSS decreased by 0.81 and FPS decreased by 3.9 compared with YOLOX-S. Through experiments, it can be seen that the SPPH module has a great influence on the accuracy of bird's nest detection, the EIOU loss function has the greatest influence on the convergence rate of loss, and the SPPH module has the greatest influence on the detection speed. However, the overall reduced FPS of the model has no influence on real-time detection, which is within the allowable range.

Figure 10 compares the effects of different algorithms on bird's nest detection. In the first column of pictures, the bird's nest was less blocked by transmission towers, and YOLOX-SCHE had the best detection effect, with a confidence level of 0.4 higher than YOLOX-S. In the second column of pictures, the bird's nest was seriously blocked by transmission towers, and YOLOX-SCHE could still detect the bird's nest accurately, with a confidence level 0.5 higher than YOLOX-S. In the third column of pictures, affected by illumination and shape, YOLOX-S failed to accurately detect the position of bird's nest, with a confidence of 0.44, which was below the set standard confidence. YOLOX-SCHE accurately detected the nest's location with a confidence of 0.67. In the following pictures, YOLOX-SCHE had the highest confidence, and the target box for detecting the bird's nest was more suitable for the actual size, which could effectively detect the bird's nest in different parts of the transmission tower.

(a)

(b)



(c)

(d)

4. **Conclusion.** In view of the complex background of UAV aerial images, the bird's nest target is small and variable in shape, and the bird's nest is easily blocked by transmission towers. In this paper, the bird's nest detection method of transmission tower based on

(e)

FIGURE 10. Comparison of detection results in different experiments. (a) is the input image. (b) is the image after YOLOX-S detection. (c) is the image after YOLOX-SC detection. (d) is the image after YOLOX-SCH detection. (e) is the image after YOLOX-SCHE detection

improved YOLOX-S is proposed, which mainly makes the following three improvements: (a) CBAM attention mechanism is added before feature fusion to reduce irrelevant information interference. (b) SPPH structure is used to enhance the feature extraction ability of the backbone network and provide more accurate initial feature information for feature fusion. (c) EIOU loss function mainly solves the IOU loss function two boxes do not intersect leading to the problem of being difficult to assess and speed up the model convergence. In the experiment, it is verified that the accuracy of the model can be improved to different degrees after the sequential addition of the three improved methods, which have good complementarity and compatibility. As the above three improvement measures do not significantly improve network computing parameters, the overall detection speed doesn't change much. Compared with the original YOLOX-S algorithm, it can improve recognition accuracy and maintain recognition efficiency while carrying out bird's nest detection for transmission towers in aerial images.

Although our model achieves better results, there are still difficulties in the detection of small-sized nests with special shapes. If the occlusion is very serious, it will also cause more difficulties in detection. In future work, we will continue to optimize the model and add a large number of bird's nest datasets. Enhance the learning and generalization capabilities of the network to improve the above issues.

REFERENCES

[1] M. Ju and C. D. Yoo, "Detection of bird's nest in real time based on relation with electric pole using deep neural network," in *2019 34th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)*. IEEE, 2019, pp. 1-4.

[2] J. Xu, J. Han, Z. G. Tong and Y. X. Wang, "Method for detecting bird's nest on tower based on UAV image," *Computer Engineering and Applications*, vol. 53, no. 6, pp. 231-235, 2017.

[3] L. Shi, Y. R. Chen, G. D. Fang, K. Y. Chen and H. Zhang, "Comprehensive identification method of bird's nest on transmission line," *Energy Reports*, vol. 8, no. 6, pp. 742-753, 2022.

[4] P. N. Huu, D. N. Tien and K. N. Manh, "Action recognition application using artificial intelligence for smart social surveillance system," *Journal of Information Hiding and Multimedia Signal Processing*, vol. 13, no. 1, pp. 1-11, 2022.

[5] S. Q. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137-1139, 2017.

[6] J. F. Dai, Y. Li, K. He and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*. Curran Associates Inc, 2016, pp. 379-387.
[7] K. He, G. Gkioxari, P. Dollar and R. Girshick, "Mask R-CNN," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 386-397, 2020.
[8] Z. W. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 6154-6162.
[9] Q. Hu and L. Zhai, "RGB-D image multi-target detection method based on 3D DSF R-CNN," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 33, no. 8, pp. 1-15, 2019.
[10] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu and A. C. Berg, "SSD: Single shot multibox detector," in *European Conference on Computer Vision*. Springer, 2016, pp. 21-37.
[11] T. Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollar, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318-327, 2020.
[12] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You only look once: unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016, pp. 779-788.
[13] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018. [Online]. Available: https://doi.org/10.48550/arXiv.1804.02767
[14] D. G. Xu, L. Wang and F. Li, "Review of typical object detection algorithms for deep learning," *Computer Engineering and Applications*, vol. 57, no. 8, pp. 10-25, 2021.
[15] X. Wu, P. Yuan, Q. Peng, C. W. Ngo and J. Y. He "Detection of bird nests in overhead catenary system images for high-speed rail," *Pattern Recognition*, vol. 51, no. 3, pp. 242-254, 2016.
[16] F. Li, J. B. Xin, T. Chen, Y. L. Li and Y. Zhang, "An automatic detection method of bird's nest on transmission line tower based on Faster RCNN," *IEEE Access*, vol. 8, no. 1, pp. 164214-164221, 2020.
[17] K. Zhao, J. D. Li, J. Huang, J. Y. Zheng and S. J. Zhang, "Bird nest detection in distribution network based on Cascade R-CNN," *Zhejiang Electric Power*, vol. 40, no. 3, pp. 73-78, 2021.
[18] Y. C. Zhong, S. Y. Sun, S. Lu, Z. Y. Luo, Y. L. Xiong and H. Q. He, "Recognition of bird's nest on transmission tower in aerial image of high-volage power line by YOLOv3 algorithm," *Journal of Guangdong University of Technology*, vol. 37, no. 3, pp. 42-48, 2020.
[19] D. Satheeswari, L. Shanmugam and N. M. J. Swaroopan, "Recognition of bird's nest in high voltage power line using SSD," in *2022 First International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)*. IEEE, 2022, pp. 1-7.
[20] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2107.08430
[21] A. Bochkovskiy, C. Wang and H. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," 2020. [Online]. Available: https://doi.org/10.48550/arXiv.2004.10934
[22] H. Zhang, M. Cisse, Y. N. Dauphin and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," 2018. [Online]. Available: https://doi.org/10.48550/arXiv.1710.09412
[23] C. Wang, H. M. Liao, Y. Wu, P. Chen, J. Hsieh and I. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2020, pp. 1571-1580.
[24] S. Hao, X. Zhang, X. Ma, S. Y. Sun, H. Wen, J. L. Wang and Q. L. Bai, "Foreign object detection in coal mine conveyor belt based on CBAM-YOLOv5," *Journal of China Coal Society*, 2022. [Online]. Available: https://doi.org/10.13225/j.cnki.jccs.2021.1644
[25] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan and S. Belongie, "Feature pyramid networks for object detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 936-944.
[26] S. Liu, L. Qi, H. Qin, J. Shi and J. Jia, "Path aggregation network for instance segmentation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 8759-8768.
[27] Z. Y. Cheng, G. Q. Xu, B. Xu and J. Luo, "Construction of anchor-free hand gesture detection network under complex background," *Computer Engineering and Design*, vol. 42, no. 6, pp. 1742-1748, 2021.
[28] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016, pp. 770-778.