Recognition of Bird Nests on Power Transmission Lines in Aerial Images Based on Improved YOLOv4

Zhaoyun Zhang* and Guanfeng He

School of Electronic Engineering and Intelligence, Dongguan University of Technology, Dongguan, China

Bird nests on transmission line towers pose a serious threat to the safe operation of power systems. Exploring an effective method to detect bird nests taken by drone inspection is crucial. However, the images taken by drones have problems such as drastic changes in the size of the object, occlusion of the object, and inconsistency in the characteristics of the object in relation to the background. The original YOLOv4 model has difficulty solving these problems. Therefore, this article improves the original YOLOv4 model by adding a Swin transformer block to its backbone network, fusing the attention mechanism into the neck of the original model, implementing classification and regression tasks for head decoupling, and using an anchor-free frame strategy and the SimOTA sample allocation method. The improved model was trained and tested on a bird nest dataset, and the detection accuracy reached 88%. Finally, the method was compared and evaluated against Faster R-CNN, RetinaNet, SSD, and the original YOLOv4, four of the other mainstream object detection models. The results showed that the accuracy obtained by the algorithm was better than the other models; the algorithm could effectively detect difficult objects such as multiple angles, occlusions, and small objects, and the detection speed could meet the real-time requirements.

Keywords: bird nest, YOLO, object detection, power inspection, smart grid

INTRODUCTION

Birds are indispensable and important members of nature. They play a vital role in the global ecosystem and directly affect the human health, economic development, and food production, as well as millions of other species. However, their nesting behavior has always been one of the main sources of transmission line faults (Li et al., 2020a). Bird nests on the transmission lines can easily cause the lines to trip or insulators to be broken down, causing major losses to the operation of the distribution network. To ensure the safe operation of the transmission network and reduce the potential safety hazards to the transmission lines caused by bird activities, it is necessary to monitor the behavior of bird nests on the overhead line towers.

Traditional manual inspection methods are labor-intensive and have low inspection efficiency and personal safety. In some dangerous terrains, inspections cannot even be carried out (Ding et al., 2021). To this end, power grid companies have introduced new technologies at a large scale in recent years, using robots, drones, and helicopters to perform fault inspection on the overhead lines (Dai et al., 2020). With the popularization of a new generation of inspection methods, the difficulty of field inspections has been greatly reduced, but massive amounts of visualized data have been produced. Faced with a large number of inspection images and videos, the use of naked eyes to detect the

FIGURE 1 | Architecture of the improved YOLOv4 model.

FIGURE 2 | Ghost module structure.

inspection images is not efficient and has greatly increased the burden on the staff (Li et al., 2020b). As a result, currently, the analysis and the processing of power inspection images mainly involve uploading all the inspection data to the backend server through network transmission, using the powerful computing power of the server to store and perform object detection. Chen et al. performed CenterNet-based bird nest detection of overhead lines in power grids and introduced the anchor-free mechanism to overcome the disadvantages of excessive preselection frame calculations in the existing object detection algorithms. Wang et al., (2019) used the Faster R-CNN algorithm to detect a bird nest on a tower, with a multiscale algorithm alleviating the difficult detection problem under a complex background. Ding et al., (2021) proposed a dual-scale bird nest detection algorithm based on YOLOv3, which not only took into account the accuracy and efficiency of the detection algorithm but also had strong antinoise performance and improved the robustness of the detection algorithm. Liu et al., (2020) proposed an algorithm based on RetinaNet, which improved the detection accuracy of small-sized bird nests by increasing the number of feature layers and expanding the range of the network's receptive field. Liu et al., (2020) improved the spatial pyramid module of YOLOv4 to reduce the loss of object information due to pooling. It also

improved the loss function to enhance the model's ability to distinguish similar objects.

However, most of the existing object detection algorithms have been designed for images of natural scenes. Because of the randomness of the position of bird nests, it is impossible to shoot at a fixed position, in contrast to the fixed position of the insulator on the overhead line. The direct application of an existing detection algorithm to deal with the object detection task in the UAV capture scene has the following main problems (Li et al., 2021). First, drastic changes in the flying height of the drones drastically change the proportion of the detected bird nests. Second, the overhead line images taken by the drones contain high-density objects (such as towers), which can cause the bird nests to be obscured (Tian et al., 2021). Third, due to the large coverage area, images taken by the drones always contain confusing geographic elements, forming the illusion that bird nests blend with the background (Zhang et al., 2021). The abovementioned three problems make the automatic detection of bird nests based on drone photography very challenging. In terms of the problems of the original YOLOv4 model for detecting small and dense objects, it requires a strong prior knowledge to set the size of the preselection box (Wang et al., 2021); the Swin transformer block (Wang et al., 2020) module is added to the backbone network, and the Swin transformer block is used. The module makes up for the lack of global information extraction capabilities of the convolutional network and further improves the feature extraction capabilities of the model. The Swin transformer block module and convolutional block attention module (CBAM) are also added to the neck to further strengthen the neck multiscale feature map spatial information and semantic information fusion ability and convert the information output by the backbone network into a feature map with a more contextual information input to the detection head. A decoupled head operation is performed in the detection head part, and the classification and regression tasks are processed separately. This enables the model to achieve better accuracy and convergence speed during the classification and regression and introduces an

FIGURE 3 | Two successive Swin transformer blocks.

anchor-free frame mechanism, eliminating the need for design. It can obtain the detection accuracy comparable to an anchor-based mechanism and uses SimOTA for positive and negative sample matching, while determining the priori box step, which greatly alleviates the problem of the positive and negative sample mismatch. The improved model architecture is shown in **Figure 1**. Our contributions are listed as follows:

- We integrated the ghost module into YOLOv4, which can significantly decrease the amount of model parameters and computation.
- We integrated the decoupled head into YOLOv4, which can slightly accelerate the training speed.
- We integrated the CBAM into YOLOv4, which can help the network to find the region of interest in images that have a large region coverage.
- We integrated the Swin transformer block into YOLOv4, which can accurately localize the objects in high-density scenes.

BACKBONE

The backbone network is used as a feature extraction network to extract the image information for generating a feature map and then detecting the location and the category of the object. The existing object detection models often use classification networks

with powerful feature extraction capabilities such as ResNet, MobileNet, EfficientNet, and Darknet as the backbone network, but the parameters of the backbone network need to be fine-tuned according to specific detection tasks. Based on the original YOLOv4 backbone network CSPDarknet, the CSP module is replaced with a ghost module, the original convolutional downsampling layer is replaced with a ghost convolution, and finally, the Swin transformer block is added to the end of the backbone network.

Ghost Module

Based on the YOLOv4 backbone network CSPDarknet-53, the original backbone network of CSPDarknet is optimized. First, three ghost layer modules and four ghost convolutional downsampling modules are added to the backbone network. Each ghost layer module is based on the residual module of Darknet53 and is produced by the improvement of the structure of CSPNet (Han et al., 2020). It consists of three ghost convolution modules and n ghost bottleneck modules that are superimposed. A ghost module first uses a small number of convolution kernels to extract the features of the input feature map, then further performs cheaper linear change operations on this part of the feature map, and generates the final feature map through the splicing operation, as shown in **Figure 2**.

Dosovitskiy et al., (2020) pointed out that rich feature information could be captured by stacking the convolutional layers containing redundant information, which would be conducive to a more comprehensive understanding of the data by the network. Therefore, the rich feature information is extracted through the conventional convolution operation, and the redundant feature information is generated by a cheaper linear transformation operation. This can effectively reduce the computing resources required by the model, and the design is simple and easy to implement, allowing for a plug-and-play execution.

Swin Transformer Block

Inspired by the vision transformer (ViT) (Liu et al., 2021), the last CSP module in the original YOLOv4 version of CSPDarknet is replaced with a transformer encoder block. Compared with the original CSP module, the transformer encoder block can capture global information and richer contextual information. However, the introduction of the transformer module greatly increases the number of calculations and parameters of the model. In contrast to the direct multi-head self-attention of the global feature map in the ViT, the Swin transformer uses the concept of window multi-head self-attention (W-MSA) to divide the feature map into multiple disjoint windows. As shown in **Figure 3**. Multi-head self-attention is performed only within each window. It can reduce the amount of calculation significantly, especially when the shallow feature map is large. Although this reduces the amount of calculation, it also isolates the transfer of information between different windows, thereby giving rise to the shifted window multi-head self-attention (SW-MSA) operation in the Swin transformer. Since W-MSA and SW-MSA are used in pairs, this solves the problem of information exchange between different windows.
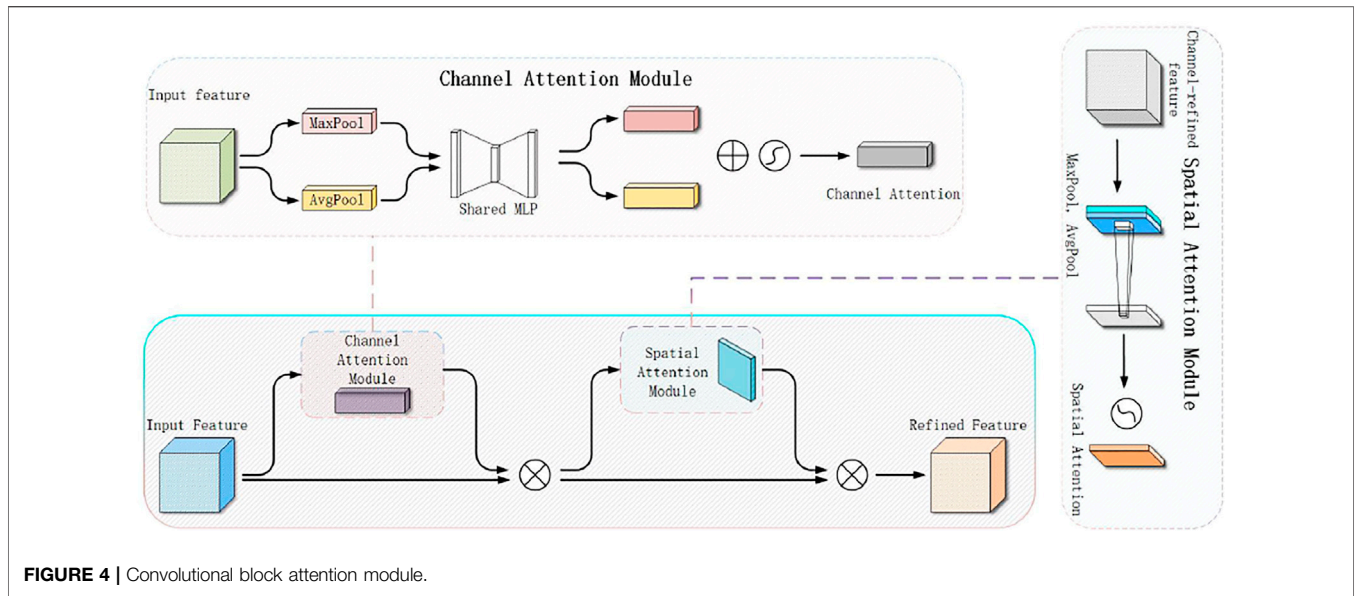
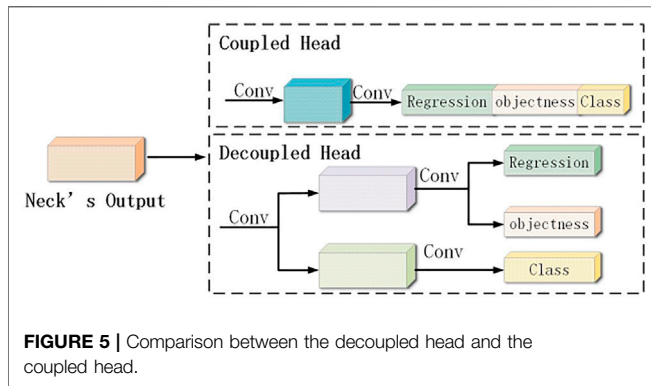**FIGURE 4 |** Convolutional block attention module.

FIGURE 5 | Comparison between the decoupled head and the coupled head.

The Swin transformer block increases the ability to capture different local information while using the self-attention mechanism to explore the potential of feature representation. In addition, it performs better on the high-density occluded objects. Due to the trade-off among the amount of calculation, the number of parameters, and the accuracy, the Swin transformer block is applied only to the neck and the end of the backbone network based on YOLOv4. Because the resolution of the feature map at the end of the network is low, applying the Swin transformer block to the low-resolution feature map can reduce the expensive calculation and storage costs, while allowing the model to pay more attention to the extraction of bird nest features.

NECK

To make better use of the features extracted by the backbone network, the feature maps of different stages extracted by the backbone are reprocessed and used rationally. The following three main improvements are made: 1) the ghost module

replaces the CSPLayer module; 2) the convolution block attention module is added to the upsampling and downsampling path of the neck; and 3) the Swin transformer block is added to the output of the neck module. The ghost module is the same as that in the backbone network. Therefore, no special introduction is given in this section.

CBAM

The convolutional block attention module is simple but effective. It is a lightweight module that can be integrated and is used for plug-and-play execution in any convolutional neural network and thus can be trained end-to-end. Given an intermediate feature map, the convolutional block attention module sequentially infers the attention map along the two independent dimensions of channel and space and then multiplies the attention map with the input feature map to perform the adaptive feature optimization. The structure of the convolution block attention module is shown in **Figure 4**. According to the experiment in (Woo et al., 2018), after integrating the convolution block attention module into different models on different classification and detection datasets, the performance of the model is greatly improved. This proves the effectiveness of this module. According to the images taken by drones, large coverage areas always contain complex background elements. Using the convolution block attention module, we can extract the attention regions, help the model resist chaotic information, and focus on the detection of bird nests.

Swin Transformer Block

The fusion of multiscale features has always been difficult in object detection. The image contains rich visual semantic information in space and scale. This semantic information is continuously obtained through the convolution operations, and only the information within the size range of the convolution
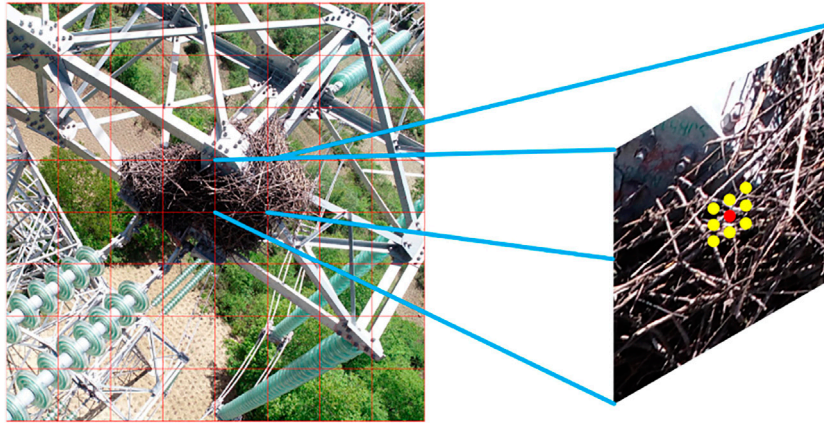
FIGURE 6 | Positive sample candidate area.

TABLE 1 | SimOTA algorithm.

SimOTA algorithm process

1. Identify positive sample candidate areas.
2. The regression loss and category loss of each sample and the GT are calculated to generate the cost matrix M.
3. According to the sum of the IoU values of the top ten samples of the current GT, the number of positive samples k is determined.
4. In M, the first K samples with the minimum loss are taken for each GT as positive samples.

kernel can be extracted each time. To obtain global information, it is necessary to stack multiple convolutional layers. In this regard, some researchers have proposed a nonlocal neural network (NLN), which matches the nonlocal information in the convolutional layer. Wang et al., (2018) believed that the NLN does not cross-scale and only extracts information across space, unable to capture the nonlocal context of objects at different scales. Because the neck itself is a multiscale feature fusion module, the cross-scale feature interaction can effectively locate and identify the local details of semantics. Zhang et al., (2020) believed that the existing methods cannot achieve the cross-scale feature interaction and have proposed the feature pyramid transformer (FPT) to make full use of the mutual fusion of the cross-space and cross-scale features. The advantage of introducing the Swin transformer block module is to connect the low-resolution, high-semantic information and low-level features of the high-resolution, low-semantic information output from the different stages of the backbone network, from top to bottom and vice versa. Therefore, the features at different scales' output by the neck have rich semantic information.

HEAD

As a classification network, the backbone network cannot complete the positioning task, and the head is responsible for detecting the location and the category of the object through the feature map extracted from the backbone network. Detectors are generally divided into two categories: one-stage object detectors and two-stage object detectors. The two-stage detector has long been the dominant method in the field of object detection, and the most representative one is the RCNN series. In contrast to the two-stage detector, the one-stage detector can predict the boundary regression and the object category at the same time. The single-stage detector has obvious advantages in speed but has a lower accuracy. To improve the capabilities to detect and position small objects, based on the original YOLOv4 detection head, the classification and regression tasks are decoupled, the anchor-free mechanism is introduced, and the SimOTA sample allocation algorithm is used.

Decoupled Head

In object detection, the conflict between classification tasks and regression tasks is a long-standing problem. However, the coupled heads for classification and positioning are widely used in most one- and two-stage detectors. Song et al., (2020) proposed that the positioning and classification tasks of object detection have a spatial misalignment problem. In other words, the two tasks have different focuses and places of interest, and the classification pays more attention to the similarity between the extracted features and the existing categories, while the positioning pays more attention to the coordinate position of the real frame. Therefore, to make the
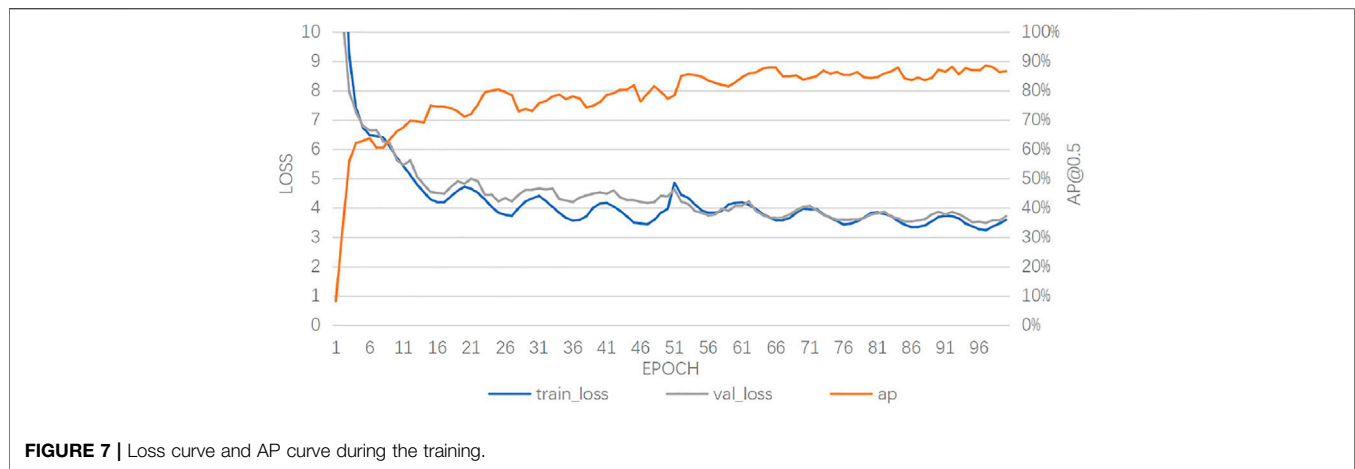
FIGURE 7 | Loss curve and AP curve during the training.

TABLE 2 | Test results of the different detection algorithms.

Model	Precision (%)	Recall (%)	AP@0.5 (%)	Params (M)	Flops (G)	Fps (rtx3090)
Faster R-CNN	52.80	77.09	74.49	28	473	25f/s
RetinaNet	93.94	67.93	76.19	36	81	49f/s
SSD	93.47	74.10	83.32	23	136	67f/s
YOLOv4	86.34	83.07	87.31	63	70	43f/s
Proposed	92.11	83.67	88.56	30	40	43f/s

detection head more efficient, the method of decoupling the head is introduced, that is, the detection head is changed from one branch to three branches using a convolution operation, as shown in **Figure 5**. Decoupling the head leads to an improved detection accuracy of the model and a faster convergence speed. Although the 1 × 1 convolutional dimensionality reduction operation is performed on the feature map output by the neck, decoupling the head does slightly slow the inference speed of the model. In general, however, head decoupling has more advantages than disadvantages.

Anchor-Free Mechanism

Due to the variation in the shooting angle and the distance of drones and the size of bird nests, the sizes of bird nests in different images are quite varied, and artificially set anchor frames cannot always match the real frames in the dataset. The anchor frame mechanism needs to set the size and the scale parameters in advance, which requires a strong prior knowledge. To address this problem, the traditional YOLO series uses the K-means method to cluster the dataset and calculate the size of the anchor frame. The use of K-means does increase the detection performance, but the generated anchor frame can only apply to a specific dataset, and it increases the complexity of the detection head and the number of potential object predictions. In addition, the

traditional anchor-free frame method uses only the center position of each grid as a positive sample (the red dot in **Figure 6**), ignoring other high-quality prediction samples away from the center. Optimizing those high-quality prediction samples is not only beneficial to the convergence speed but also alleviates the imbalance of the positive and negative samples during the training (Ge et al., 2021). In addition, for discontinuous objects such as bird nests, only a positive sample is taken at the center of the object, which is likely to cause errors and affect the final detection performance. Therefore, multiple samples of the 3 × 3 area in the object center are selected as the positive sample candidates (the red and yellow dots in **Figure 6**).

SimOTA

The traditional object detection algorithm allocates the positive and negative samples according to the IoU of the anchor frame and the real frame. However, the division of the positive and negative samples under different sizes, shapes, and occlusion conditions should be different, and the context information also needs to be considered. An excellent sample matching algorithm can effectively solve the dense object detection problem and optimize the detection effect when there are extreme proportions of objects or imbalanced positive samples of extreme size objects (Tian et al., 2019). Therefore, SimOTA treats sample matching as an optimal

FIGURE 8 | Test results of different detection models.
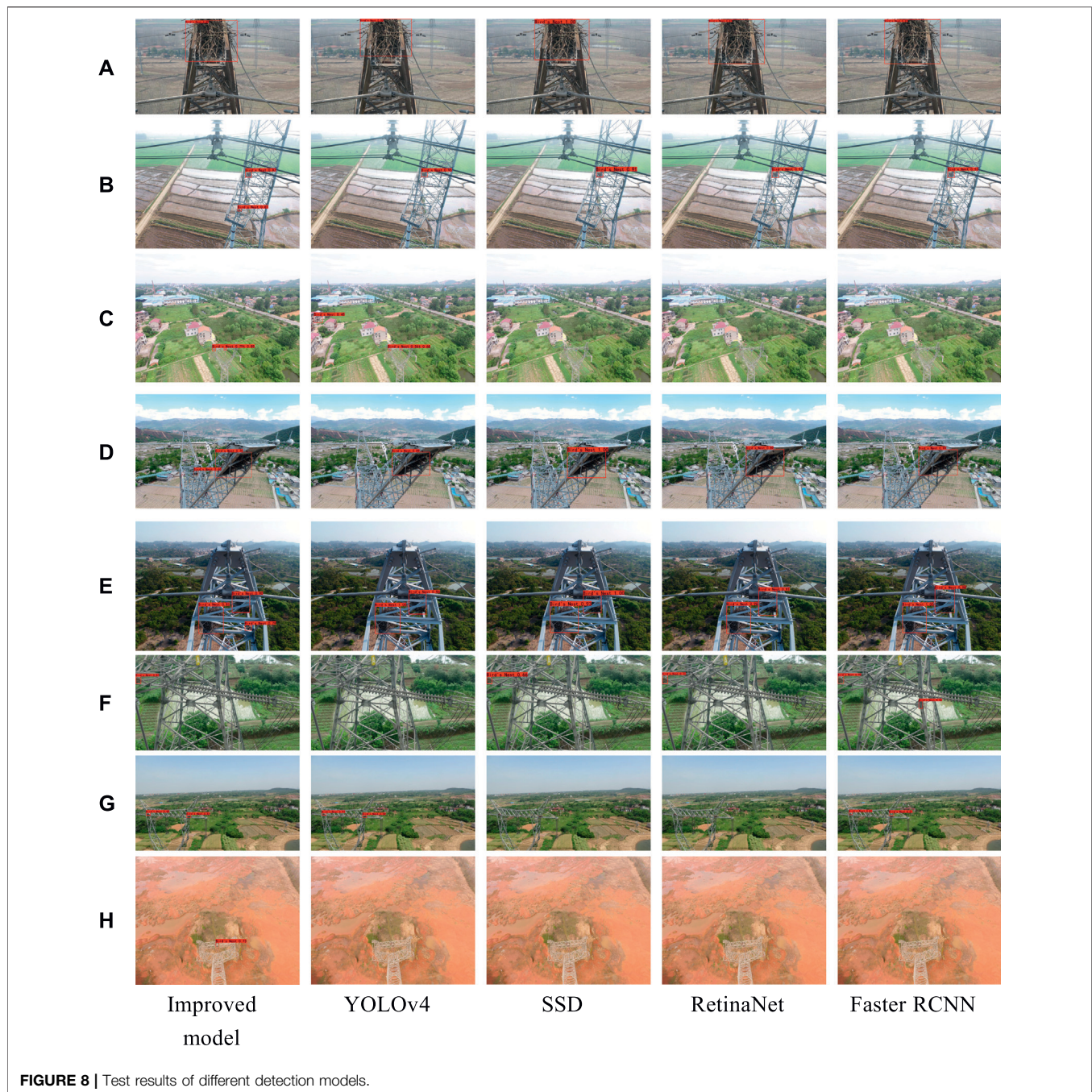
transportation problem in a linear programming problem. The specific process is shown in **Table 1**.

EXPERIMENT

The experiment selected a real UAV line patrol dataset containing bird nests. The proposed model and the four mainstream models of YOLOv4, Faster R-CNN, SSD, and RetinaNet were trained, tested, compared, and evaluated on the selected dataset.

Description

The bird nest dataset comes from the data obtained by a power grid company conducting live-line inspections with unmanned aerial vehicles on its overhead lines. The dataset for this study had a total of 4,514 bird nest pictures, and the original dataset was divided into non-test and test sets at a

TABLE 3 | Ablation study result.

Model	AP@0.5 (%)	Params (M)	Flops (G)	Fps (rtx3090)
YOLOv4	87.31	63	70	43f/s
YOLOv4+ghost	82.54	35	39	65f/s
YOLOv4+ghost + decoupled head	80.22	29	42	51f/s
YOLOv4+ghost + decoupled head + CBAM	84.54	29	42	47f/s
Proposal (previous + Swin transformer)	88.56	30	40	43f/s

ratio of 9:1. The non-test set was divided 9:1 into training and validation sets. The training set, the validation set, and the test set included 3,655, 407, and 452 images, respectively. Because the bird nest detection model requires a large number of data samples to train the network, this study performed random mosaic data enhancement and a series of random processing of color gamut and size before inputting each picture in the original dataset.

Training and testing on the experimental data were carried out on the same deep learning server with the Ubuntu 18.04 operating system, an Intel Xeon W-2245 CPU, a single GeForce RTX 3090 24 GB GPU, and 64-GB DDR4 RAM. The training and testing were implemented using the PyTorch 1.8.0 framework, and the detection effect of a single picture or a video was visualized through the OpenCV tool library.

Analysis

At the beginning of the training, the weights of the model were randomly initialized, and the coco dataset was used to pre-train the model and to let the model learn the ability to extract features. This can extract the potential features or common structures between the original problem dataset and the object dataset, thereby accelerating the training and improving the performance of the model. The training process was optimized by batch normalization. Each batch has trained 16 samples, and each iteration has trained 228 batches. Since the bird nest dataset was relatively small, only 100 epochs were trained. The initial learning rate of the model was set to the cosine annealing learning rate, the Adam optimizer was used, the weight attenuation was 0.0005, and the size of the input image was 640×640. The parameters of the backbone network were not updated in the first 50 epochs, and the entire parameters of the network were updated in the rest of the 50 epochs. The training results are shown in **Figure 7**.

Result

To verify the effectiveness of the improved model, Faster R-CNN, RetinaNet, SSD, YOLOv4, and the improved model were evaluated experimentally on the same bird nest dataset, and the experimental results were compared (all the models were trained under the same conditions). Evaluation indicators included the calculation accuracy (precision), average precision (AP) with a detection threshold of 0.5 (IoU), model parameters (Params), model calculations (Flops), and detection speed (FPS). Among them, AP is an

important indicator for evaluating the detection effect, which can be obtained by calculating the area enclosed by the curves for the accuracy rate P and recall rate R. **Table 2** shows the test results of the five models. Among them, the improved model had an AP value of 88.56% and a detection speed of 43 f/s when the threshold was 0.5. The analysis is as follows: the improved model was superior to the other models in detection accuracy and recall rate. Among the models, RetinaNet achieved the highest accuracy due to the lower regression rate. However, overall, the detection effect of the improved model was better than the other types of detection. In addition, while the introduction of head decoupling into the improved model resulted in a slower detection speed than that of the SSD algorithm and RetinaNet, it still met the requirements of real-time detection.

Figure 8 shows the detection results of the improved model and the four other mainstream detection models on the test set. The test set included a small-sized bird nest object, multiple bird nest objects in a single image, a bird nest being severely obscured by a tower, and a bird nest blending with the background. These are the common situations for a variety of power inspections. The improved model maintained a good detection performance under the above mentioned conditions. Other models had good detection capabilities for bird nests (as shown in **Figure 8A**). However, the following situations were also observed. 1) Insufficient detection capabilities for the small objects: as shown in **Figures 8B,E**, the SSD algorithm, RetinaNet algorithm, Faster R-CNN algorithm, and YOLOv4 algorithm failed to detect the third bird nest at the bottom right. Each of the above mentioned algorithms also missed the detections in **Figures 8C,D,F,G**. 2) They were prone to false detection: as shown in **Figure 8C**, YOLOv4 recognized the bamboo frame on the left as a bird nest. As shown in **Figure 8F**, Faster R-CNN recognized the tower on the right as a bird nest. 3) The detection capability for complex backgrounds was insufficient. As shown in **Figure 8H**, all the four mainstream models missed the detection. The aforementioned results show that the improved model can maintain the detection accuracy and robustness in different scenarios and is suitable for the power inspection of the UAV aerial photography.

Ablation Study

The importance of each proposed component was analyzed on the test set. The impact of each component is listed in **Table 3**.

CONCLUSION

The classic object detection algorithm is not suitable for the detection of aerial images, and it has disadvantages such as poor detection accuracy, high rates of missed detection, and excessive model scale. Based on YOLOv4, improvements were made from the backbone network, neck, and detection head. The ghost module, the CBAM module, the transform module, the anchor-free mechanism, and SimOTA were added based on YOLOv4 to form an improved bird nest detection algorithm of drone aerial photography. The improved model was tested on a bird nest dataset from the aerial photographs of an electric power inspection drone. The experiments showed that compared with the other mainstream algorithms, the improved algorithm had advantages in the recall rate and simultaneously obtained a higher accuracy rate. Therefore, the improved model can help power inspectors to obtain a better experience in UAV power inspection. Although the proposed algorithm presently meets the requirements for real-time bird nest detection, the detection speed still needs to be improved compared with the SSD algorithm. In the future, by reducing the scale of the weight parameters of the network, the model will be lightened, and the detection speed will be improved.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

ZZ and GH contributed to conception and design of the study. ZZ performed the statistical analysis. GH wrote the first draft of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

REFERENCES

Chen, Z. Y., Li, Y., Wang, D. H., and Zhang, S-J. (2021). A Device Anomaly Detection Method Based on CenterNet [J]. *Comp. Appl.* 41 (S1), 304.

Dai, Z., Yi, J., Zhang, Y., Zhou, B., and He, L. (2020). Fast and Accurate Cable Detection Using CNN. *Appl. Intell.* 50 (12), 4688–4707. doi:10.1007/s10489-020-01746-9

Ding, J., Huang, L., and Zhu, D. (2021). High Tower as the Bird's Nest Detection Oriented Double Scale YOLOv3 Network Study [J]. *J. Xi 'an Polytechnic Univ.* (02), 253–260. doi:10.19322/j.carolcarroll

Dosovitskiy, A., Beyer, L., and Kolesnikov, A. (2020). *An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale[J]*. arXiv preprint arXiv:2010.11929.

Ge, Z., Liu, S., and Wang, F. (2021). *Yolox: Exceeding yolo Series in 2021.* [J]arXiv preprint arXiv:2107.08430.

Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., and Xu, C. (2020). "Ghostnet: More Features from Cheap Operations[C]," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1580–1589.

Li, F., Xin, J., Chen, T., Xin, L., Wei, Z., Li, Y., et al. (2020). An Automatic Detection Method of Bird's Nest on Transmission Line Tower Based on Faster_RCNN. *IEEE Access* 8, 164214–164221. doi:10.1109/access.2020.3022419

Li, J., Yan, D., Luan, K., Li, Z., and Liang, H. (2020). Deep Learning-Based Bird's Nest Detection on Transmission Lines Using UAV Imagery. *Appl. Sci.* 10 (18), 6147. doi:10.3390/app10186147

Li, X., Li, Z., and Wang, H. (2021). Unmanned Aerial Vehicle for Transmission Line Inspection: Status, Standardization, and Perspectives[J]. *Front. Energ. Res.* 9, 336. doi:10.3389/fenrg.2021.713634

Liu, G. W., Zhang, C. X., and Liu, B. (2020). An Improved RetinaNet Model for Catenary Bird Nest Detection. *Data Acquisition Process.* 35 (03), 563–571. doi:10.16337/j.1004-9037.2020.03.018

Liu, Z., Lin, Y., and Cao, Y. (2021). *Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows[J]*. arXiv preprint arXiv:2103.14030.

Song, G., Liu, Y., and Wang, X. (2020). "Revisiting the Sibling Head in Object Detector[C]," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 11563–11572.

Tian, G., Liu, J., Zhao, H., and Yang, W. (2021). Small Object Detection via Dual Inspection Mechanism for UAV Visual Images[J]. *Applied Intelligence*, 1–14.

Tian, Z., Shen, C., Chen, H., and He, T. (2019). "Fcos: Fully Convolutional One-Stage Object Detection[C]," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 9627–9636.

Wang, C. Y., Bochkovskiy, A., and Liao, H. Y. M. (2021). "Scaled-yolov4: Scaling Cross Stage Partial Network[C]," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 13029–13038.

Wang, C. Y., Liao, H. Y. M., Wu, Y. H., Chen, P. Y., Hsieh, J. W., and Yeh, I. H. (2020). "CSPNet: A New Backbone that Can Enhance Learning Capability of CNN[C]," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (Seattle, USA, 390–391.

Wang, J. W., Luo, H. B., and Yu, P. F. (2019). Multi-Scale Bird Nest Detection for High Pressure Tower Based on Faster R-Cnn [J]. *Beijing Jiaotong Univ.* 43 (05), 37. doi:10.11860/j.issn.1673-0291.20180168

Wang, X., Girshick, R., Gupta, A., and He, K. (2018). "Non-Local Neural Networks [C]," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 7794–7803.

Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. (2018). "Cbam: Convolutional Block Attention Module[C]," in Proceedings of the European Conference on Computer Vision (ECCV) (Cham: Springer), 3–19. doi:10.1007/978-3-030-01234-2_1

Zhang, D., Zhang, H., Tang, J., Wang, M., Hua, X., and Sun, Q. (2020). "Feature Pyramid Transformer[C]," in European Conference on Computer Vision (Cham: Springer), 323–339.

Zhang, X., Wan, T., Wu, Z., and Du, B. (2021). *Real-Time Detector Design for Small Targets Based on Bi-Channel Feature Fusion Mechanism[J]. Applied Intelligence*, 1–10.