The Visual Computer (2024) 40:3003–3021
https://doi.org/10.1007/s00371-023-03004-8

ORIGINAL ARTICLE

Foreign object detection for transmission lines based on Swin Transformer V2 and YOLOX

Chaoli Tang [1] · Huiyuan Dong[1] · Yourui Huang[1,2] · Tao Han[1] · Mingshuai Fang[1] · Jiahao Fu[1]

Accepted: 24 June 2023 / Published online: 18 July 2023

Abstract

Suspended foreign objects on transmission lines will shorten the discharge distance, easily leading to phase-to-ground or phase-to-phase short circuits, which induces outage accidents. Foreign objects are small and difficult to identify, resulting in low detection accuracy. An improved foreign object detection method based on Swin Transformer V2 and YOLOX (ST2Rep–YOLOX) is proposed. First, the feature extraction layer ST2CSP constructed by Swin Transformer V2 is used in the original backbone network to extract global and local features. Secondly, hybrid spatial pyramid pooling (HSPP) is designed to enlarge the receptive field and retain more feature information. Then, Re-param VGG block (RepVGGBlock) is introduced to replace all 3×3 convolutions in the network to deepen the network and improve feature extraction capabilities. Finally, experiments are carried out on the transmission lines foreign object image dataset, which was obtained using unmanned aerial vehicles (UAVs). The experimental results show that the average accuracy of the ST2Rep–YOLOX method can reach 96.7%, which is 4.4% higher than that of YOLOX. The accuracy of the nest, kite, and balloon increased by 9.3%, 15.4%, and 9.6%, and the recall increased by 6.5%, 9.4%, and 2.5%, respectively. This method has high detection accuracy, which provides an important reference for foreign object detection in transmission lines.

Keywords Transmission lines · Foreign object detection · YOLOX · Swin Transformer V2 · RepVGGBlock

1 Introduction

Power is transmitted by transmission lines. It is particularly crucial to ensure the safe and reliable operation of transmission lines to improve the quality of transmission and the safety of the power grid [1, 2]. Therefore, the failure of transmission lines should be detected and eliminated in time to avoid serious safety accidents and economic losses. With the growth of the economy and the large-scale development of transmission lines, the investment needed for line maintenance in the late stage also increases. In China's vast territory, transmission lines need to cross a variety of terrains, and the geographical environment is usually remote and complex, making power equipment inspection work difficult [3, 4]. Due to people's lives and the natural environment, foreign objects, such as nests, kites, balloons, and other foreign objects, are easily attached to transmission lines and likely to cause a phase-to-ground short circuit or a phase-to-phase short circuit, which in turn leads to regional power outages, causing serious economic losses and even endangering the safety of people around the transmission lines [5]. It poses a great threat to the safe operation of the power system. So,

✉ Huiyuan Dong
2021200743@aust.edu.cn

Chaoli Tang
chltang@mail.ustc.edu.cn

Yourui Huang
hyr628@163.com

Tao Han
than@aust.edu.cn

Mingshuai Fang
fangms2023@163.com

Jiahao Fu
2020200703@aust.edu.cn

[1] School of Electrical & Information Engineering, Anhui University of Science and Technology, Huainan 232001, China

[2] School of Electrical and Electronic Engineering, West Anhui University, Lu'an 237012, China

it is necessary to increase the intensity of the inspection of transmission lines and clean up the hidden dangers of foreign bodies to prevent foreign object attachment caused by line accidents.

The inspection of transmission lines in China is mainly based on manual inspection, which is inefficient, laborious, and frequently consumes plenty of resources [6]. UAVs are now widely used to inspect transmission lines and other electrical equipment due to their advantages of portability, simple operation, and rapid response [7, 8]. They not only reduce the labor intensity and protect the personal safety of workers, but they also help to improve the efficiency of transmission line inspection and avoid many line outage accidents [9]. The main methods used for foreign object detection are traditional artificial-based detection methods and deep learning-based detection methods.

Traditional artificial detection methods mainly rely on the texture structure, appearance and transform domain decomposition of the target to preprocess the image. The commonly used feature discrimination methods: oriented gradient histogram (HOG) [10] and scale invariant feature transform (SIFT) [11], and then design feature detection methods and descriptors to extract the features of given images through a lot of experience. The process is complicated and can only extract shallow features. The method based on image morphology [12, 13] generally uses a filter to remove noise, then applies Otsu (maximum variance between classes) to segment the background and foreground of the image, and finally extracts power lines to identify foreign objects using Hough transform[14]. Hazgui et al. [15] propose a genetic programming (GP)-based method that combines the two well-known features of histograms of oriented gradients and local binary patterns to simultaneously perform patch detection, feature extraction and image classification. Lu et al. [16] designed a method based on the cascade classifier and combined features for power transmission line inspection. The image is described by multi-angle features and is recognized by the cascade classifier. These manual detection methods rely on experts to design special feature extraction methods and have poor generalization performance, which may result in a decrease in detection accuracy when the background conditions change. At the same time, when the UAVs capture the image, the color and shape of the image will change due to the influence of illumination, shooting distance, and angle [17]. In using these methods for detection, the results are susceptible to interference from the surrounding environment, and the detection accuracy is affected.

The deep learning-based detection algorithm can obtain deeper feature representations based on the learning of a large number of samples, which are more efficient and accurate for the expression of datasets. The extracted abstract features are more robust and have better generalization ability. Target detection algorithms can be divided into two-stage target detection algorithms, such as the region convolution neural network (R-CNN) series, and one-stage target detection algorithms, such as the You Only Look Once (YOLO) series [18–20]. Zhao et al. [21] improved the Faster RCNN model as the detection algorithm of foreign objects in transmission lines. Finally, the detection model can recognize the fault types with a mean average precision (mAP) of 90.8% for glass insulators and 91.7% for composite insulators. Zhang et al. [22] designed multi-view Faster R-CNN based on tensor decomposition. Compared with object detection methods YOLOv3, SSD, and Faster R-CNN, the improved Faster R-CNN model had lower miss probability and higher detection accuracy. Xu et al. [23] proposed an efficient substation foreign object detection network that consists of a moving target area extraction network and a classification network. The results showed that the model performed better than Fast R-CNN and Mask R-CNN. Sarkar et al. [24] equipped Raspberry Pi with a test image as an input to detect an insulator's health status using YOLOv3 and used a super-resolution CNN to reconstruct a blurred image as high-resolution image. Li et al. [25] put forward a lightweight YOLOv3 model running on an embedded device to detect foreign objects of transmission lines. The improved model has a smaller model size and higher detection speed without notably reducing detection accuracy. Qiu et al. [26] designed a lightweight YOLOv4 model with embedded dual attention mechanism (YOLOv4-EDAM) to detect foreign objects from visible images. Cui et al. [27] discussed the problems of video object detection and introduced a framework named TF-Blender which contains temporal relation, feature adjustment, and feature blender modules to solve the problem of feature degrading in the video frames. Su et al. [28] designed EpNet to generate region proposals at the edge of the image to reduce complex backgrounds; in the pre-training, massive synthetic data are applied to alleviate the problem of data shortage and enhance the performance of foreign object detection. Wang et al. [29] devised deep nearest centroids (DNC) which conducts nonparametric, case-based reasoning. It performs better on image classification and boosts pixel recognition with improved transparency, using various backbone network architectures. Wang et al. [30] put forward a newer version of YOLO—YOLOv7—which was faster and more accurate than others. At present, most of the detection methods have good accuracy, but in cases of a complex image background and tiny target, there are still problems with target false detection and missed detection, and the speed of real-time detection is not high.

To solve the problems of insufficient accuracy and lack of robustness in the process of foreign object detection, this paper propounds a foreign object detection method for transmission lines based on YOLOX [31]. The main contributions of the paper are as follows:

- We propose a framework called ST2Rep–YOLOX, which can better capture global and local feature information, improve detection accuracy and reduce false detection and missed detection of foreign objects in transmission lines.
- In ST2Rep–YOLOX, we devise ST2CSP module to extract global and local information in the backbone network, and a re-param module called RepVGGBlock [32] to reinforce the expression ability of the model in the training phase and merge models to reduce the model parameters in the inference phase.
- Our proposed module contains the designed HSPP module, which expands the receptive field and retains more information compared with the original spatial pyramid pooling (SPP) [33] module. In the neck, dual-branch structure feature pyramid network (FPN) and path aggregation network (PAN), and the designed ST2CSP module are used to efficiently aggregate semantic features and location features, making full use of each feature layer.
- Our module of foreign object detection is simple and effective. Compared to the latest algorithm YOLOV7 [30], we can obtain better performance in mAP and apply fewer parameters.

2 Materials and methods

2.1 Original YOLOX detection model

YOLOX based on the YOLOv3 algorithm improves many aspects, such as data augmentation, predictive branch decoupling, anchor-free frame, and Simple Optimal Transport Assignment (SimOTA) label allocation [34]. Compared to previous YOLO series algorithms, YOLOX has advantages in detection accuracy and speed. As a derivative version of the YOLOX model, YOLOX-s has a simple structure and fewer parameters, which are easy to deploy. With comprehensive consideration, YOLOX-s was chosen as the benchmark model. Figure 1 shows the YOLOX-S network structure diagram.

The YOLOX-s network structure is divided into four parts: input, backbone network, neck network, and prediction output. Before training, YOLOX-s employs Mosaic and Mixup data augmentations to preprocess the input image. The network uses the Cross Stage Partial Darknet (CSPDarknet) to extract features. The neck network adopts the structure of FPN. From top to bottom, the high-level feature information is transmitted and fused by upsampling to obtain a feature map for prediction. The prediction output decouples the classification and regression branches, resulting in a series of fixed prediction sets, which includes four-coordinate information of the prediction target box, one target confidence score, and N kinds of prediction scores.

2.2 Improved YOLOX detection method

The improved algorithm inherits the advantages of YOLOX-s. In the transmission lines foreign object detection task, Mosaic and MixUp data augmentations were used to enrich the detection background information and strengthen the generalization ability of the model. Based on the YOLOX-s network architecture, we introduced the Swin Transformer V2 and add RepVGGBlock to the network in order to extract deeper features. The HSPP perceptual field module was designed to obtain more detailed feature information and improve detection accuracy. The network structure of the improved YOLOX-s is shown in Fig. 2, which is described in detail below.

2.2.1 Swim Transformer V2 model

The original YOLOX network model uses CSPDarknet based on the convolution structure as the backbone network for feature extraction. In the process of extracting features by convolution, the size of the receptive field depends on the size of the convolution kernel. The larger the convolution kernel, the larger the range of the region. However, the increase in the convolution kernel will greatly increase the complexity of the operation. When the receptive field area is not large enough, the global feature information will be lost. The convolution structure has translation invariance and is insensitive to the global position of the information, which leads to only extracting a small part of the local information in the original data.

Swin Transformer uses an attention mechanism that takes into account global information when calculating attention. By adding location information to each patch, the receptive field is enlarged while retaining global location sensitivity to features. Swin Transformer V2 was upgraded based on V1. In Fig. 3, compared to V1 (a), Swin Transformer V2 (b) has three differences, which are highlighted in red: (1) post-normalization of model stability; (2) the dot-product-generated attention is replaced by a cosine attention calculation; (3) logarithmic interval continuous position offset replaces the original relative position offset [35].

Based on these preponderances, Swin Transformer V2 has higher large-scale visual model stability and better cross-window resolution migration model performance.

The input feature map is assumed to be F. Window-based multi-headed self-attention (W-MSA) is performed first and then normalized in the Swin Transformer module. The calcu-

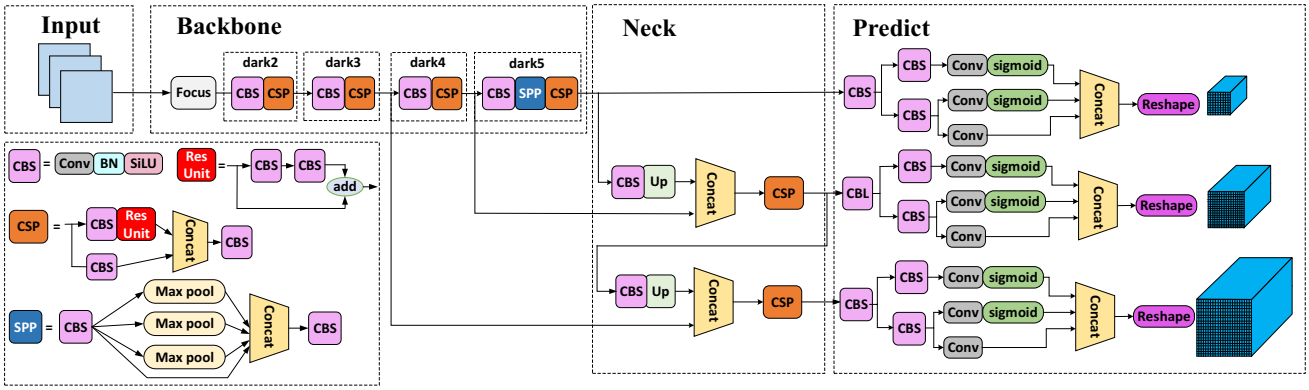


Fig. 1 The structure of the original YOLOX-s network model detection

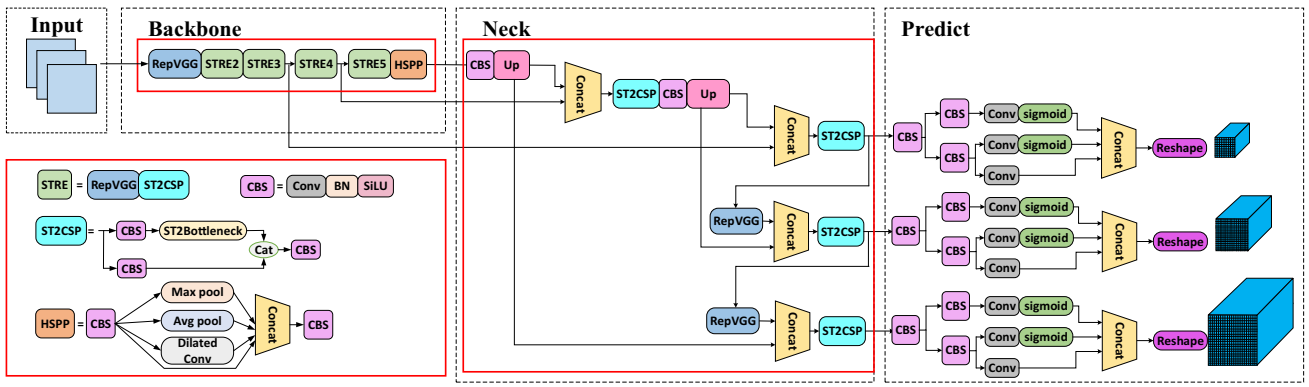


Fig. 2 The structure of improved YOLOX-s network model detection

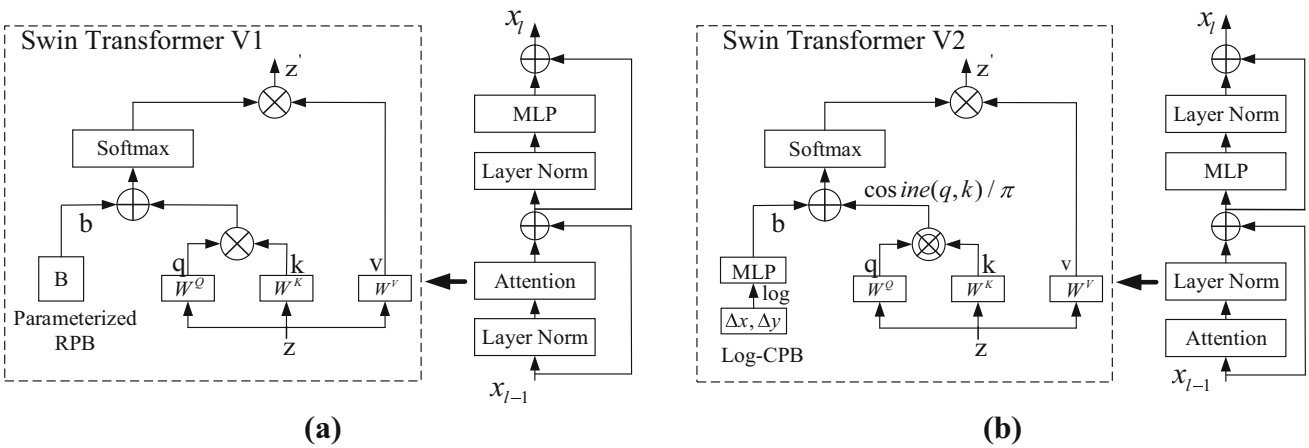


Fig. 3 **a** The structure of Swin Transformer V1; **b** the structure of Swin Transformer V2

lation process of the window-based multi-head self-attention is as follows. In the formula, Q, K, and V are the query, key, and value vectors. $B(\Delta x, \Delta y)$ represents the continuous position offsets of the relative intervals of each pixel.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{\cos(q, k)}{\tau} + B(\Delta x, \Delta y)\right) v \tag{1}$$

By introducing $B(\Delta x, \Delta y)$, the spatial position relationship between pixels is maintained, which can avoid the loss of position information of the input sequence. The computational complexity of self-attention is linearly related to the size of the input feature map, while W-MSA calculates self-attention in the divided small window, which greatly reduces the computational complexity. However, it only

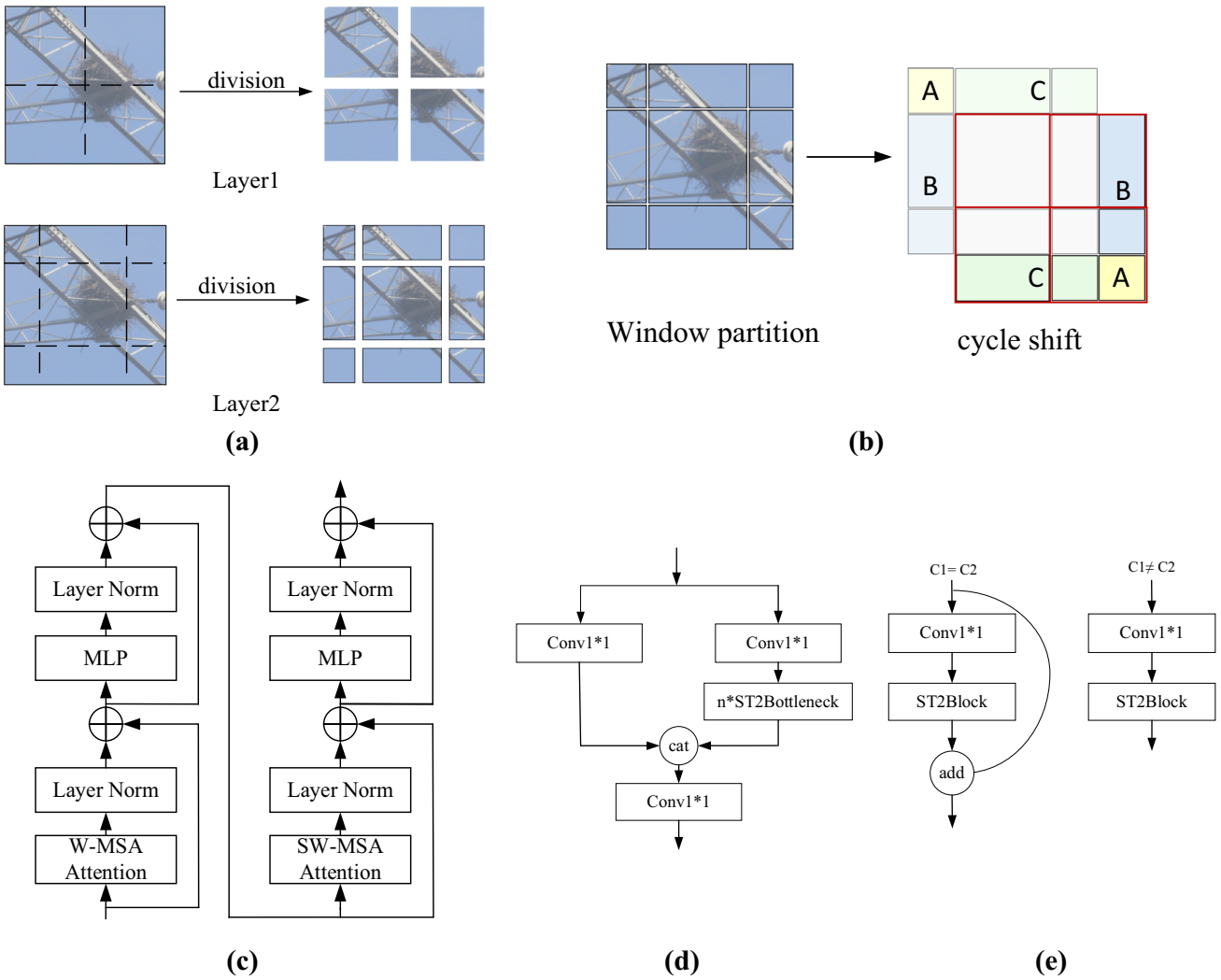


Fig. 4 **a** illustrations of W-MSA and SW-MSA; **b** the cyclic shift operation of SW-MSA; **c** The components of Swin Transformer V2 Block; **d** ST2CSP; **e** ST2Bottleneck

obtains information inside the window, and there is no information exchange between the windows, so it is impossible to obtain global features. At the time, the shifted window-based multi-headed self-attention (SW-MSA) operation is needed after the W-MSA operation.

As Fig. 4a shows, W-MSA employs the conventional window partition strategy that the feature map of 8 × 8 size is divided into 2 × 2 patches in size of 4 × 4. However, SW-MSA is obtained by shifting the patch position by 1/2 patch size, which includes 3 × 3 non-overlapping patches. The division of the shifted window introduces a connection between the adjacent non-overlapping windows of the upper layer, which greatly increases the receptive field. To keep the same number of patches, the patch of the upper left corner A, B, and C that do not satisfy the 4 × 4 scale after translation is spliced with the patch of the lower right corner in Fig. 4b. Although the number of patches seems invariant, it satisfies the information interaction outside the window, which is called cyclic shift [36]. SW-MSA enables the information interaction between different windows, enabling the network to capture more context information. The backbone network combines the advantages of a convolutional layer and Swin Transformer V2, taking into account the local information and global information, and can learn more distinguishable features. Figure 4 shows the structure of the Swin Transformer V2 Block and the illustrations of W-MSA and SW-MSA.

The ST2CSP module constructed by the Swin Transformer V2 block replaces the Cross Stage Partial (CSP) in CSPDarknet for feature extraction. Firstly, bottleneck uses the residual network Residual, which can be divided into two parts, as Fig. 4d, e shows. The backbone part is a 1 × 1 convolution and a Swin Transformer V2 component. The

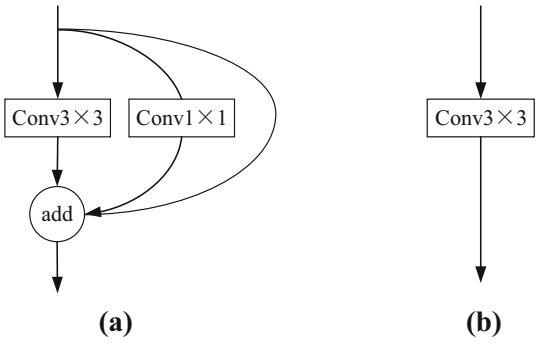


Fig. 5 **a** RepVGG in the model training; **b** RepVGG in the model deploying

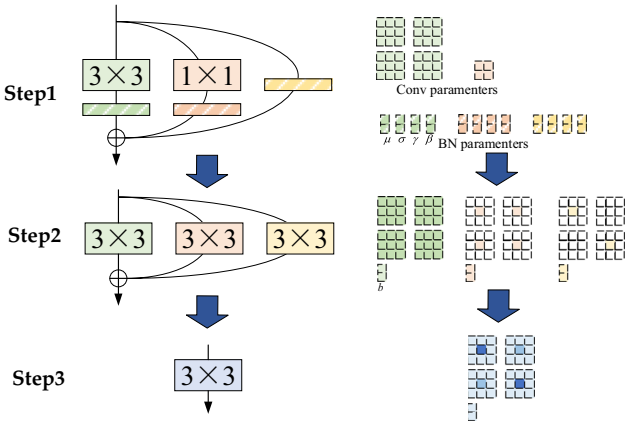


Fig. 6 RepVGG fusion process structure diagram

residual edge portion directly combines the input and output of the trunk. Secondly, the ST2CSP module is built by the CSPnet network structure, which consists of two parts: the main part performs 1 × 1 convolution and the stacking of bottleneck residual blocks; the other part, like a residual edge, is connected directly to the end by the 1 × 1 convolution. This module can improve the extraction effect of global features. At the same time, compared to the original network, the number of parameters is slightly reduced.

2.2.2 RepVGGBlock model

RepVGG is a classification network, which combines the ideas of the VGG network and the ResNet network. As Fig. 5 shows, in the model training stage, Identity and 1 × 1 convolution residual branches are added to the block of the VGG network, which is equivalent to applying the characteristics of the ResNet network to the VGG network. In the model inference phase, all network layers are converted to 3 × 3 convolution through the Optimizer fusion strategy to facilitate model deployment and acceleration.

Figure 6 shows the process of fusion [32]. Each branch is individually converted to 3 × 3 convolution, and the converted convolutions of the three branches are merged into a new 3 × 3 convolution. The fusion details in the inference stage are as follows.

Step 1: The convolution and BN layer in the residual block are fused by Eq. 2.

Step 2: Convert the fused convolutional layer to 3 × 3 convolution. For the 1 × 1Conv branch, the value in the 1 × 1 convolution kernel can be moved to the center point of the 3 × 3 convolution kernel; for the Identity branch, the branch does not change the value of the input feature map, so it can be regarded as a 3 × 3 convolution kernel with a weight value of 1. Then it is multiplied by the input feature map to maintain the original value.

Step 3: Merge 3 × 3 convolution in the residual branch. By superimposing the weight W and bias B of all branches, a merged 3 × 3 convolution network layer is obtained.

$$W_i' = \frac{\gamma_i}{\sigma_i} W_i \quad b_i' = -\frac{\mu_i \gamma_i}{\sigma_i} + \beta_i \tag{2}$$

In the formula, W_i is the parameter of the convolution layer before conversion. μ_i is the mean value of the BN layer. δ_i is the variance of the BN layer. γ_i and β_i represent the scale factor and offset factor of the BN layer. W' and b' represent the weight and bias of the convolution after fusion, respectively.

RepVGGBlock is used to replace all 3 × 3 convolutions in backbone and neck. Compared to the original YOLOX, the model combined with RepVGGBlock can be regarded as a large-scale multi-branch model due to the residual structure in the training process, which adds the information obtained from different branches to strengthen the extraction of feature information. In the reasoning process, the multi-branch model is equivalently converted into a single-path model, which makes the reasoning speed fast.

2.2.3 Hybrid spatial pyramid pooling

The receptive field in the shallow feature map is small, which is not conducive to large target detection, while the receptive field in the deep feature map is large, which is not conducive to small target detection. Therefore, the original YOLOX model employs the SPP module to obtain feature maps with different receptive field sizes, so that the detection network will adapt to different sizes of targets. The SPP module consists of multi-scale sliding cores (1 × 1, 5 × 5, 9 × 9, 13 × 13) for max pooling. The four max pooling operations use stride = 1, padding, and the structure is displayed in Fig. 7.

In deep learning, in addition to downsampling the feature map through max pooling, average pooling can be adopted as well. The max pooling focuses on the maximum value of

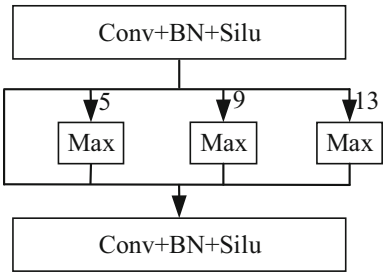


Fig. 7 The structure of SPPBottleneck

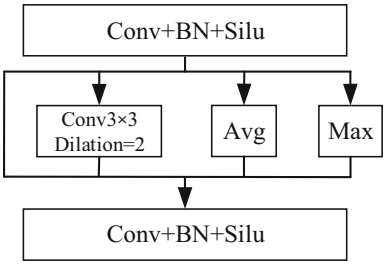


Fig. 8 The structure of HSPPBottleneck

each neighborhood, which can reduce the feature information and thus decrease the amount of calculation. Average pooling concentrates on the average of each neighborhood [37]. When dealing with images whose targets are similar to the background, average pooling can retain more target information and have a better processing effect.

To solve the problems of false detection and missed detection caused by the similarity between the target and the background in the foreign object images, the SPP module was retrofitted in this study. The original 9×9 max pooling was replaced by the average pooling. When the foreign object was similar to the background, more features and information about the target and background could be reserved. At the same time, the 3×3, dilation $= 2$ dilated convolution was used instead of a 5×5 max pooling to enlarge the receptive field without losing the resolution. The structure is shown in Fig. 8. Hybrid spatial pyramid pooling (HSPP) is designed to retain more information in scale fusion, accurately locate the target, and reduce the missed detection and error rate.

2.2.4 Loss function

As shown in Fig. 9, the prediction layer of YOLOX-s changes the YOLO head part to the Decoupled Head structure. The regression and classification are divided into two parts and combined when predicting, which reinforces the convergence speed and accuracy of the algorithm. Reg is the position information of the prediction box, including the center coordinates, width, and height information of the prediction box; obj is the object information in the prediction box, indicating

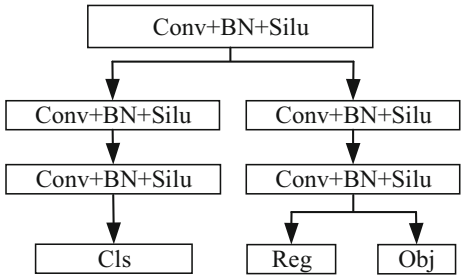


Fig. 9 The decoupled head structure

that the box contains confidential information about the existence probability of the object to be detected; cls represents the prediction box classification information.

Corresponding to the network prediction feature vector, the loss is also divided into three parts, regression loss Loss_{reg}, confidence loss Loss_{obj}, and category loss Loss_{cls}. Loss_{reg} is the IoU loss calculation, which is the ratio of the intersection area of the prediction box P and the target box G. Loss_{obj} is the binary cross entropy of the target category score obtained by multiplying the predicted probability t of the category by the predicted probability p of the confidence level.

$$\text{Loss}_{\text{reg}} = -\log(\text{IoU}) = -\log\left(\frac{P \cap G}{P \cup G}\right) \tag{3}$$

$$\text{Loss}_{\text{cls}} = -\sum_{i=1}^{n} (t_i \times \log(p_i) + (1 - t_i) \times \log(1 - p_i)) \tag{4}$$

The original YOLOX-s network uses the binary cross-entropy loss function BCELoss as confidence loss. Although it solves the problem of imbalance between positive and negative samples, it does not distinguish between easy-to-classify and difficult-to-classify samples. Aiming at the imbalance of sample classification difficulty, Focal Loss is used to modifying the CE Loss by adding the category weight and sample difficulty weight adjustment factor [38]. It is a dynamically scaled cross-entropy loss. Through a dynamic scaling factor, the weight of easily distinguishable samples in the training process can be dynamically reduced so that the center of gravity can be quickly focused on those difficult-to-distinguish samples.

$$\text{CE}(p_t) = -\log(p_t) \tag{5}$$

$$\text{FL}(p_t) = -\alpha_t (1 - p_t)^{\gamma} \log(p_t) \quad \alpha_t = \begin{cases} \alpha & \text{if } y = 1 \\ 1 - \alpha & \text{otherwise} \end{cases} \tag{6}$$

In Formula (6), the $(1 - p_t)^{\gamma}$ modulation factor is added to reduce the loss contribution of the separable samples; αt

is the weight factor, which is used to adjust the proportion between positive and negative sample losses. Positive samples use α, and negative samples use $1-\alpha$. Both γ and α have corresponding value ranges, and they interact. With the increase in γ, α should be slightly lower; setting $\gamma = 2$ and $\alpha = 0.25$ is best, as shown in reference [38].

3 Experiment and dataset

First, the pre-trained image data were obtained with the help of the UAV and then transmitted to the local PC for training and subsequent testing using the improved model. The experimental environment was as follows: Intel (R) Core (TM) i9-10900K CPU processor, 64 G memory, NVIDIA RTX3060 12G graphics cards, Windows 64-bit operating system, the deep neural network was built on the PyTorch deep learning framework, and the compiler was Pycharm.

3.1 Experimental dataset preparation

The training process of the network model needs the support of a large amount of data, and the number of datasets can affect the implementation effect of the network model to a certain extent. We mainly studied the detection of foreign objects in transmission lines, collected the image data of foreign objects in transmission lines, and then expanded the dataset utilizing rotation, translation, brightness adjustment, and contrast. Then the dataset was divided into a training set, verification set, and test set. The dataset used in this experiment is non-public and contained three foreign objects named nest, kite, and balloon. The numbers of original nest, kite, and balloon datasets were 1560, 135, and 95, respectively. Among the three types, more kite and balloon images were needed. After employing various methods of data augmentation, 135 kite images were expanded to 824 and the 95 balloon images were expanded to 618. Finally, there were 3002 datasets. As Fig. 10 shows, three types of data augmentation operations are mainly used for expanding balloon and kite images in the paper which are translation, rotation and adjusted brightness.

In this study, LabelImg was used to label the acquired foreign object detection dataset, and the label names were nest, kite, and balloon. Before the training, the numbers of the training set, validation set, and test set were 7, 2, and 1, respectively. Among them, the training set consisted of the data sample for model fitting. The validation set was a separate sample set during the model training process, which was used to adjust the model hyperparameters and preliminarily evaluate the model's ability. The test set was employed to evaluate the final generalization ability of the model. Table 1 shows the numbers of the three types of datasets.

3.2 Experimental hyperparameter setting

In order to achieve a better training effect, this study adopted the following training strategies:

(1) *Batch processing* To alleviate the limitation of hardware on the training, the training samples were processed in batches. A batch is the number of pictures to be processed by the iterative model each time. The size of the batch has an impact on the gradient descent speed of the network to a certain extent. Choosing a larger batch will increase the gradient descent speed of the network. However, due to hardware limitations, a batch that is too large will make the memory burst and interrupt the training process. To ensure a normal training process, a batch = 8 in the training phase.

(2) *Learning rate* The learning rate (lr) is the amount of weight update during training, and it is a configurable hyperparameter used in network training, with a value range of 0.0 to 1.0. Excessive lr accelerates learning in the early stage of model training, making the model easy to approach local or global optimal solutions, but it may cause the value of the loss function to oscillate in the later stage and make it difficult to achieve a real optimal solution. If the learning rate is too small, the convergence speed of the net loss will decrease, or even may not fall into the range of suboptimal solutions [39]. Therefore, the lr was set at 0.001 in the early stage of training, and the lr was moderately reduced by cosine annealing in the later stage. The cosine annealing attenuation refers to the adjustment of the lr in the form of the cosine function, which decreases slowly, then accelerates, and then decreases slowly. The update mechanism of the lr is as follows.

$$global\_step = \min(global\_step, decay\_steps) \tag{7}$$

$$\cos ine\_decay = 0.5(1 + \cos(\pi *)) \tag{8}$$

$$decay = (1-\alpha) * \cos ine\_decay + \alpha \tag{9}$$

$$decayed\_learning\_rate = learning\_rate * decayed \tag{10}$$

In the above formulas, *learning _ rate* represents the initial lr, *global _ step* represents the global number of steps used for the decay calculations, *decay _ steps* represents the number of decay steps, and α represents the minimum learning rate.

(3) *Optimizer* Optimizer is a weight parameter-updating algorithm that makes the loss function continuously approach the global minimum in the backpropagation process of a deep learning network. In the most primitive Stochastic Gradient Descent (SGD) method, the calculation amount is too large, and it is easy to converge to the local minimum. Therefore, we chose the Adaptive Moment Estimation (Adam) gradient descent method with a stronger convergence ability and

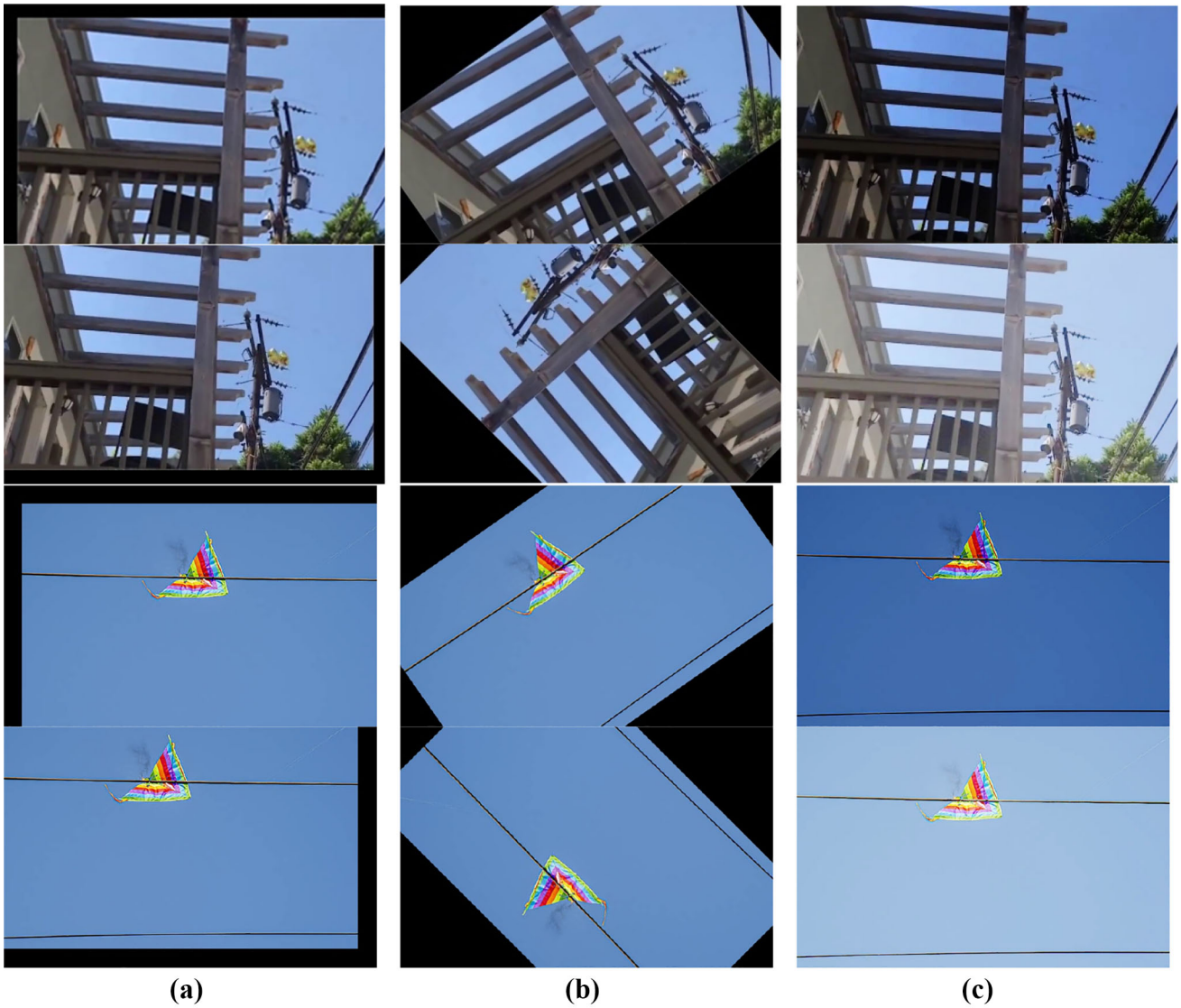


Fig. 10 Data augmentation. **a** Translation; **b** rotation; **c** adjusted brightness

Table 1 Numbers of three types of datasets

Type	Dataset				
	Original	Expand	Train	Val	Test
Nest	1560	–	1052	361	147
Kite	135	824	561	186	77
Balloon	95	618	412	129	77

higher calculation efficiency [40].

$$m_t = \beta_1 \cdot m_{t-1} + (1-\beta_1) \cdot g_t \tag{11}$$

$$v_t = \beta_2 \cdot v_{t-1} + (1-\beta_2) \cdot g_t^2 \tag{12}$$

$$\begin{cases} \theta_{t+1} = \theta_t - \dfrac{\eta}{\sqrt{\hat{v}_t} + \varepsilon} \hat{m}_t \\ \hat{m}_t = \dfrac{m_t}{1-\beta_1^t} \\ \hat{v}_t = \dfrac{v_t}{1-\beta_2^t} \end{cases} \tag{13}$$

Table 2 Hyperparameter in model training

Hyperparameter	lr	Batch	Epoch	Optimizer	IoU
Value	0.001	8	100	Adam	0.5

In the above formulas, θ_t and θ_{t+1} are the gradients at time t and $t+1$, respectively. η is the lr. m_t and $\widehat{v_t}$ represent the first and second moment estimation correction values of the gradient, respectively. β_1 and β_2 are the exponential decay rates of the first and second moment, respectively. m_t and v_t represent the first and second moments of the gradient, respectively. Table 2 displays the hyperparameter used in the model training.

3.3 Experimental evaluating indicator

To evaluate the effectiveness of the modified network more objectively, precision (P), recall (R), mean of average precision (mAP), and the number of frames per second (FPS) were selected as the evaluation indicators to detect the network performance. P was used to determine the probability of correct detection, R was used to determine whether the target in the full dataset could be found, and mAP was the mean AP value of all categories. The calculation formulas are as follows [41]:

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{14}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{15}$$

$$\mathrm{AP} = \int_0^1 p(r)\mathrm{d}r \tag{16}$$

$$\mathrm{mAP} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{AP}_i \tag{17}$$

$$\mathrm{FPS} = \frac{N}{t_{\mathrm{end}} - t_{\mathrm{start}}} \tag{18}$$

In Formulas (14) and (15), IoU is the degree of overlap between the prediction box and the ground truth. When IoU = 0.5 is set as the threshold, if it is greater than 0.5, it is True, and if it is less than 0.5, it is False. Based on the IoU threshold evaluation, TP (true positive) represents the number of positive samples correctly predicted.TP and FP represent the number of positive samples wrongly predicted, and FN is the number of undetected targets. Formula (16) is the average precision (AP), which means that the precision value obtained by the recall rate in the range of 0 to 1 is averaged. In Formula (17), APi represents the average accuracy of the i category of samples. n is the number of categories of samples in the dataset, and n = 3 in this study. mAP @ 0.5 indicates that when the IoU is 0.5, for the AP of three kinds of samples, the sum of the three is averaged to obtain the overall mAP. In Formula (18), N represents the number of images processed from the start time to the end time. The above evaluation metrics provide an objective description of the test results for the foreign object datasets on various models. The greater the values of mAP and FPS, the better the effect of the model.

4 Discussion

4.1 Comparison of receptive field modules

At the end of the backbone network, the designed HSPP was added to fuse the feature information of each scale. The HSPP receptive field module was compared with the SPP to verify the effectiveness of the HSPP module. The results are shown in Table 3.

By detecting the accuracy and recall of the three types of foreign objects and mAP@50, three methods were compared. From the data in Table 2, compared to the original SPP module, the SPP replaced by only the average pooling had a better recall rate and detection accuracy. The detection accuracy and mAP of the designed HSPP module were the highest. In summary, the designed HSPP module set different types of convolution and pooling operations to better maintain more feature information and enhance the fusion of network semantics and texture information.

4.2 Ablation experiment

To test the performance of the improved algorithm in this paper, an ablation experiment was carried out on the foreign objects dataset of the transmission lines. The new HSPP, ST2CSP module, and RepVGGBlock module were added in different combinations, and mAP@50 was selected as the performance evaluation index.

From Table 4, by comparing methods 1 and method 2, the detection accuracy of the improved algorithm employing the designed HSPP increased by 0.9%, especially nest detection improved obviously, which verifies that it is effective to use the HSPP to reduce false detection in complex background. Compared to method 1, the detection accuracy of method 3 improved by 2.1%. It shows that adding RepVGG module to deepen the network improves the accuracy of detection to a certain extent. By comparing method 1 and method 4, the P and R of all types were significantly promoted after the

Table 3 Comparison of different receptive field modules

Receptive field module	Nest		Kite		Balloon		mAP@50
	P	*R*	*P*	*R*	*P*	*R*	
SPP	0.866	0.832	0.822	0.871	0.832	0.914	0.923
Avg-SPP	0.886	0.852	0.844	0.894	0.874	0.938	0.924
HSPP	0.912	0.865	0.881	0.871	0.893	0.926	0.932

Table 4 Performance index comparison of ablation experiment

Method	HSPP	RepVGG block	ST2CSP	Nest		Kite		Balloon		mAP@50
				P	*R*	*P*	*R*	*P*	*R*	
1	×	×	×	0.866	0.832	0.822	0.871	0.832	0.914	0.923
2	√	×	×	0.912	0.865	0.881	0.871	0.893	0.926	0.932
3	×	√	×	0.893	0.865	0.876	0.918	0.915	0.926	0.944
4	×	×	√	0.923	0.852	0.988	0.976	0.927	0.938	0.956
5	√	√	×	0.920	0.884	0.932	0.965	0.928	0.951	0.958
6	×	√	√	0.959	0.897	0.976	0.941	0.950	0.938	0.959
7	√	×	√	0.915	0.938	0.965	0.976	0.938	0.983	0.962
8	√	√	√	0.959	0.897	0.976	0.965	0.928	0.939	0.967

introduction of the ST2CSP module, which indicates that the feature extraction ability of the network using the ST2CSP module was enhanced greatly. Method 5, 6 and 7 represent the experimental results of pairwise combination of three modules. As we can see, the three methods can reach better accuracy than applying only a single module. Method 6 combined RepVGG and ST2CSP performs well in accuracy of all types. Method 7 combined HSSP and ST2CSP obtains the highest recall in all types, which indicates the least missed detection. When employing three models together, a balanced result with the best mAP comes out.

From all categories of mAP in Table 4, it can be seen that the improved YOLOX greatly improved the mAP. For each type of AP, Fig. 11 shows the AP curve of each ablation experiment in detail for the three foreign objects: nest, kite, and balloon. From the following pictures, the AP of the nest increased slightly for the reason that the complex background and occlusion of the nest image made nest too difficult to detect. Comparing (a) and (d), the AP detected of the kite increased by 6%, and the balloon increased by about 2%. ST2CSP extracted more features to better detect small targets, such as balloons and kites. In column (e), (f) and (g), it can be concluded that the pairwise combinations have almost the same effect on the AP of the three categories. According to the experimental results, the improved YOLOX network in the paper was enhanced with all modules, which verifies the effectiveness of the improvement of this model.

Table 5 displays the details of the RepVGG model. We evaluate the performance from three aspects: parameters, detection accuracy and speed. Since the structure of RepVGG is deeper than that of conventional 3×3 conv, it is reasonable that adding RepVGG increases the parameters by 0.3M. With RepVGG, the detection accuracy is indeed improved. Benefiting from the fusion strategy of RepVGG, the inference speed is not slowed down but slightly improved when detecting a single image.

4.3 Contrast experiment

After 100 epochs of training, we applied the pre-trained model to predict the test set of foreign object images. As Fig. 12 shows, the image to be detected and the trained weights were loaded to obtain the predicted results.

By comparing the detection results of the YOLOX and

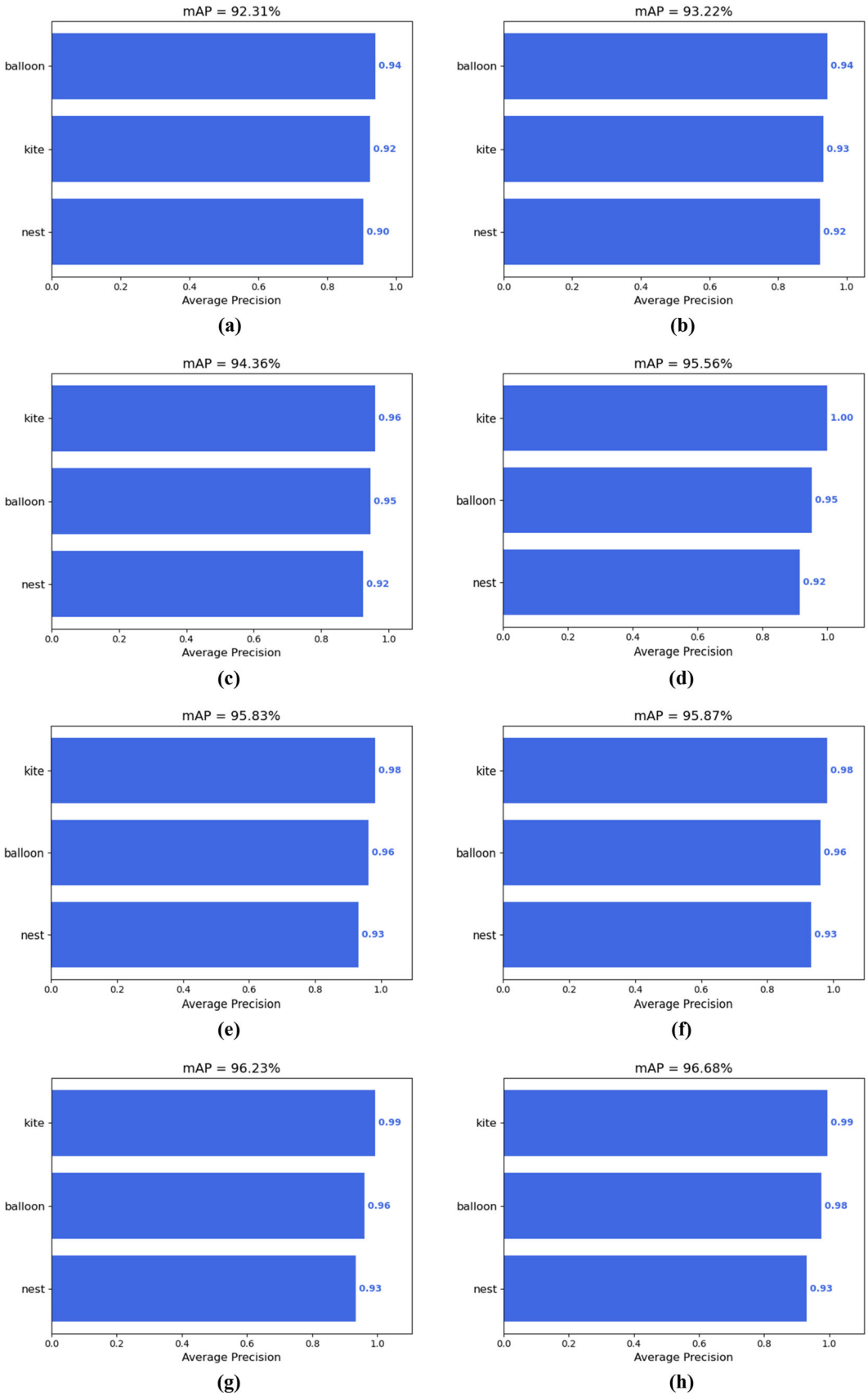


Fig. 11 **a** The *mAP* curves of method 1, **b** the *mAP* curves of method 2, **c** the *mAP* curves of method 3, **d** the *mAP* curves of method 4, **e** The *mAP* curves of method 5, **f** the *mAP* curves of method 6, **g** the *mAP* curves of method 7, **h** the *mAP* curves of method 8

Table 5 Comparison of without or with RepVGG

Method	Param	mAP@50	Time/s
WithoutRepVGG	8.9M	0.923	0.0128
WithRepVGG	9.2M	0.944	0.0125

ST2Rep–YOLOX, we can intuitively see the advantages and disadvantages of the different network models for foreign object recognition. Figure 13 shows the discrepancies between the detection results of the YOLOX network (a) and the ST2Rep–YOLOX network (b). In the figure, there are three types of foreign objects named nest, kite, and balloon. In the rectangular box of foreign object location, the target category information and the confidence level belonging to this category are displayed, respectively. The kites and balloons have smaller targets in the picture. The YOLOX network could correctly identify kites and balloons, but the confidence level was lower than that of the ST2Rep–YOLOX network. In the recognition of the nest, the images of the nest had complex backgrounds and occlusion, which made detection difficult. YOLOX had false detection. Although the improved ST2Rep–YOLOX network had low confidence, there were few missed detections and false detections. Therefore, we can conclude that the ST2Rep–YOLOX network detection effect was better.

The algorithm in the paper was compared with Faster R-CNN [42], which is a typical two-stage detection algorithm in the current target detection algorithm, YOLOv5 [43], which is representative of a one-stage detection algorithm with an anchor, the original YOLOX algorithm that is anchor-free, and the newest YOLOv7 [30] algorithm.

As can be seen in Table 6, Faster R-CNN had lowest detection accuracy and speed significantly. The performance of YOLOv5 was lower than the other algorithms, but the YOLOv5 model had the fewest parameters and was fast to detect. In the above network model, they have not achieved good balance between precision and recall rate. The latest network, YOLOv7, had the highest recall rate and the fastest detection speed for the nest and balloon. However, the number of model parameters was about four times that of YOLOX. The algorithm proposed in the paper greatly improved the detection accuracy and recall rates of the nest, kite, and balloon in the foreign object images, and the mAP was as high as 96.7%. Compared to the two-stage detection model, the detection accuracy increased by 37.2% and the detection speed increased by 18 FPS. Compared to the detection model with the anchor frame, the detection accuracy improved by 7.5%. Compared to the original model, the detection accuracy improved by 4.4%. Although the accuracy of our detection model was comparable to YOLOv7, the speed was only half of that of YOLOv7. Based on the above discussion, the results prove that the proposed method had better accuracy; however, the speed needs to be strengthened.

Figure 14 displays the detection result map under different network models, where (a), (b), and (c) are three different foreign object images to be detected. The detection results of the four network models follow. In the pictures, it can be clearly and intuitively seen that in columns (a) and (b) the Faster-RCNN is not sensitive to nest and kite, and there existed missed detection. In columns (b) and (c), it can be found that Faster-RCNN and YOLOV5 both detected the kite and balloon, but falsely detected the insulator as a balloon or kite. For the nests listed in (a), There are cases of false detection in Faster-RCNN, YOLOV5 and YOLOV7. Our proposed method was more advantageous. It can detect the nest and has high detection accuracy. For the kites listed in (b), it can be seen that the YOLOV7 model had the best detection effect. For the balloons listed in (c), although the target is small, the detection effect of the proposed method could compare favorably with that of YOLOV7. According to the network detection and image results, it can be seen that the method had higher detection accuracy.

5 Conclusions

Based on the improved YOLOX-s model integrated with Swin Transformer V2, the paper proposes a new foreign object detection algorithm for transmission lines. In this method, the improved backbone network is constructed by the multi-head self-attention of the shifted window, which

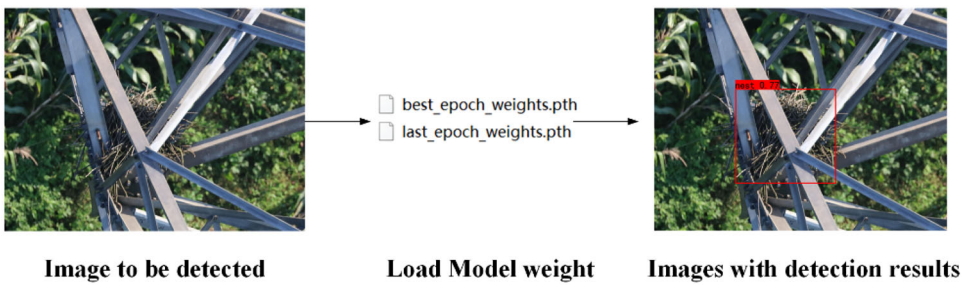


Fig. 12 The process of detection

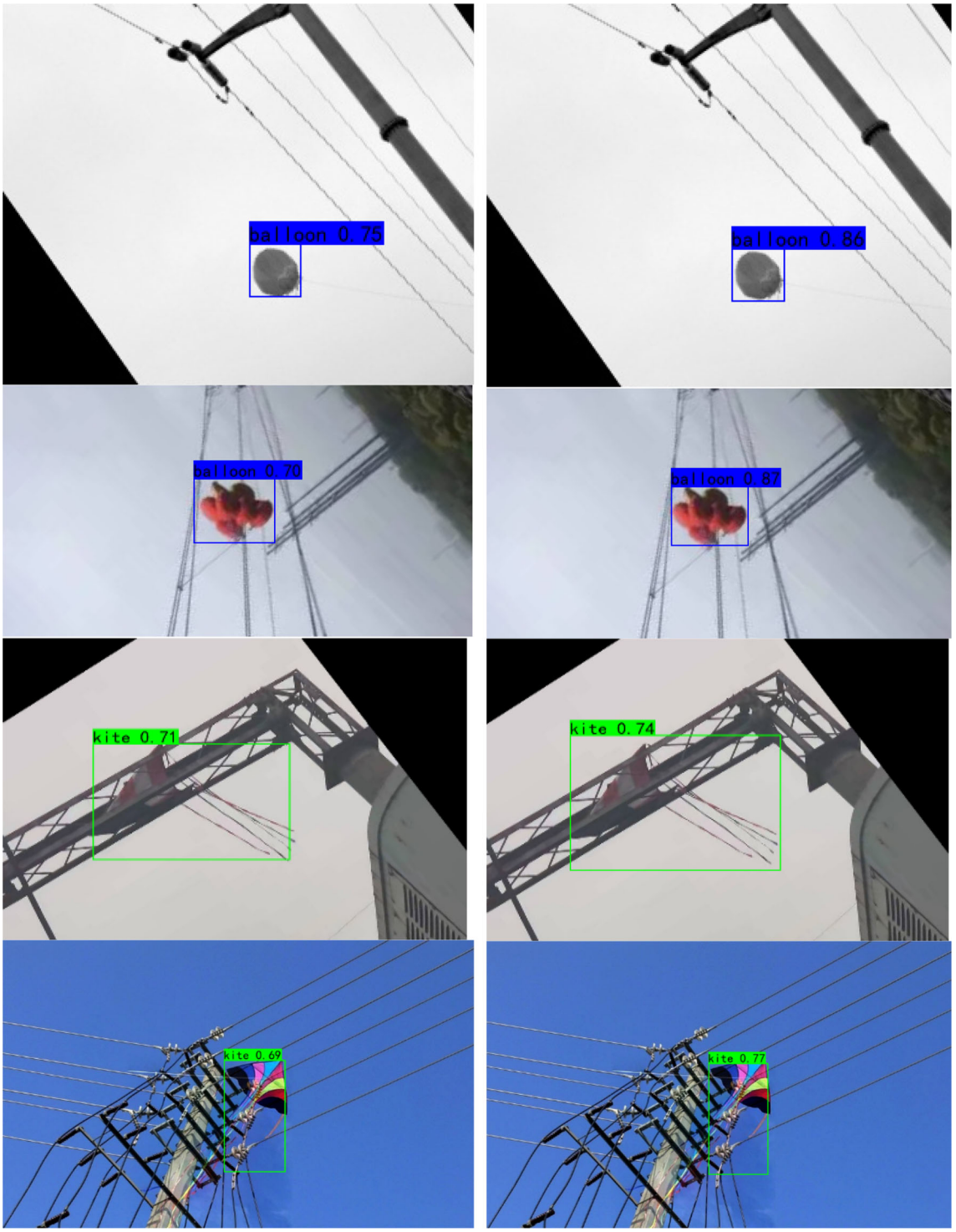


Fig. 13 **a** The test results of YOLOX. **b** The test results of improved YOLOX

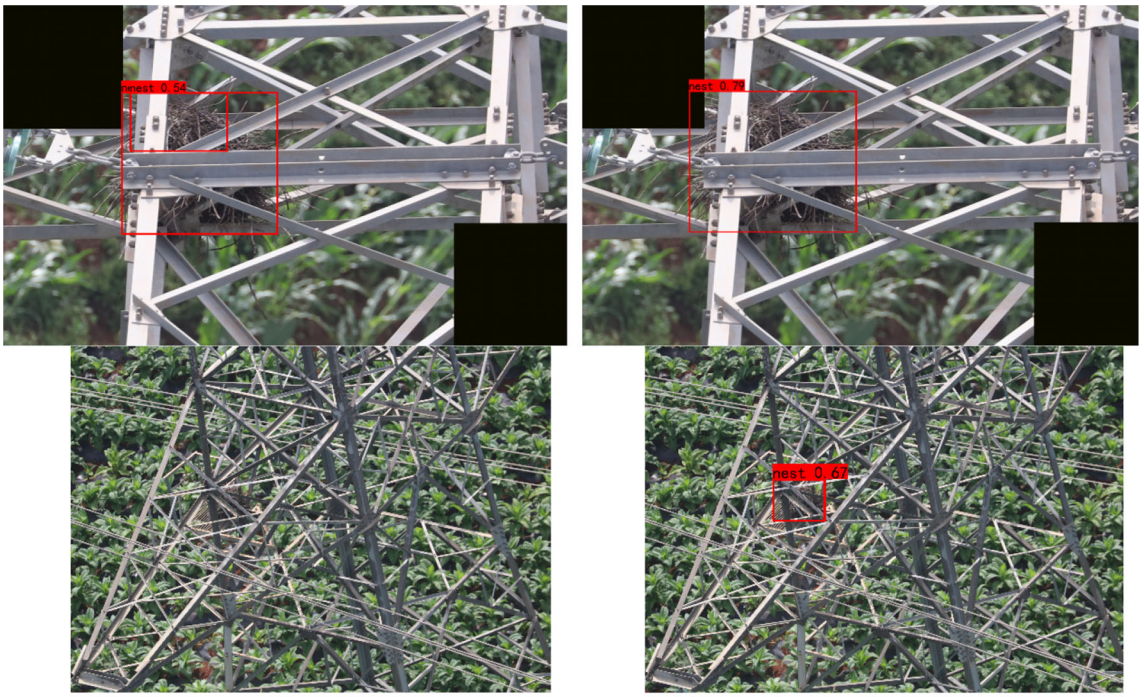


Fig. 13 continued

Table 6 Performance comparison of different models

Model	Nest		Kite		Balloon		mAP@50	FPS	Parameters
	P	*R*	*P*	*R*	*P*	*R*			
Faster-RCNN	0.369	0.691	0.284	0.718	0.454	0.728	0.595	12	137M
YOLOv5	0.867	0.835	0.910	0.547	0.949	0.902	0.892	40	**7.2M**
YOLOX	0.866	0.832	0.822	0.871	0.832	0.914	0.923	36	8.9M
Ours	**0.959**	0.897	**0.976**	**0.965**	0.928	0.939	**0.967**	30	9.6M
YOLOv7	0.930	**0.906**	0.953	0.793	**0.958**	**0.979**	0.958	**60**	36.9M

Bold values represent the optimal value of each column

can obtain more global and local information, learn more distinguishable features, and is more suitable for complex and occluded scenes. In employing the HSPP module, the receptive field is expanded and multi-scale information is fused. RepVGGBlock is adopted to further improve feature extraction ability and detection accuracy. Experiments were carried out on a transmission line foreign object target detection dataset to evaluate the algorithm. The results show that the detection accuracy mAP@50 of the improved algorithm proposed reached 96.7% and had certain advantages in accuracy and detection speed compared to the mainstream single-stage and two-stage target detection algorithms. Furthermore, the proposed model had certain disadvantages in detection speed and parameters. Our model failed to be small enough and fast enough.

In subsequent research, on the basis of ensuring the original accuracy, we will consider starting with a lightweight model so that it can be deployed in an FPGA. The proposed algorithm with an optimized detection speed in an FPGA can be applied to UAV systems or transmission lines foreign body cleaning robots for real-time online foreign object detection. Additionally, deep learning-based method can be applied to a variety of image processing tasks. Although we have achieved certain results in object detection, we need to expand the application of the algorithm in other tasks, such as classification, segmentation or image restoration.

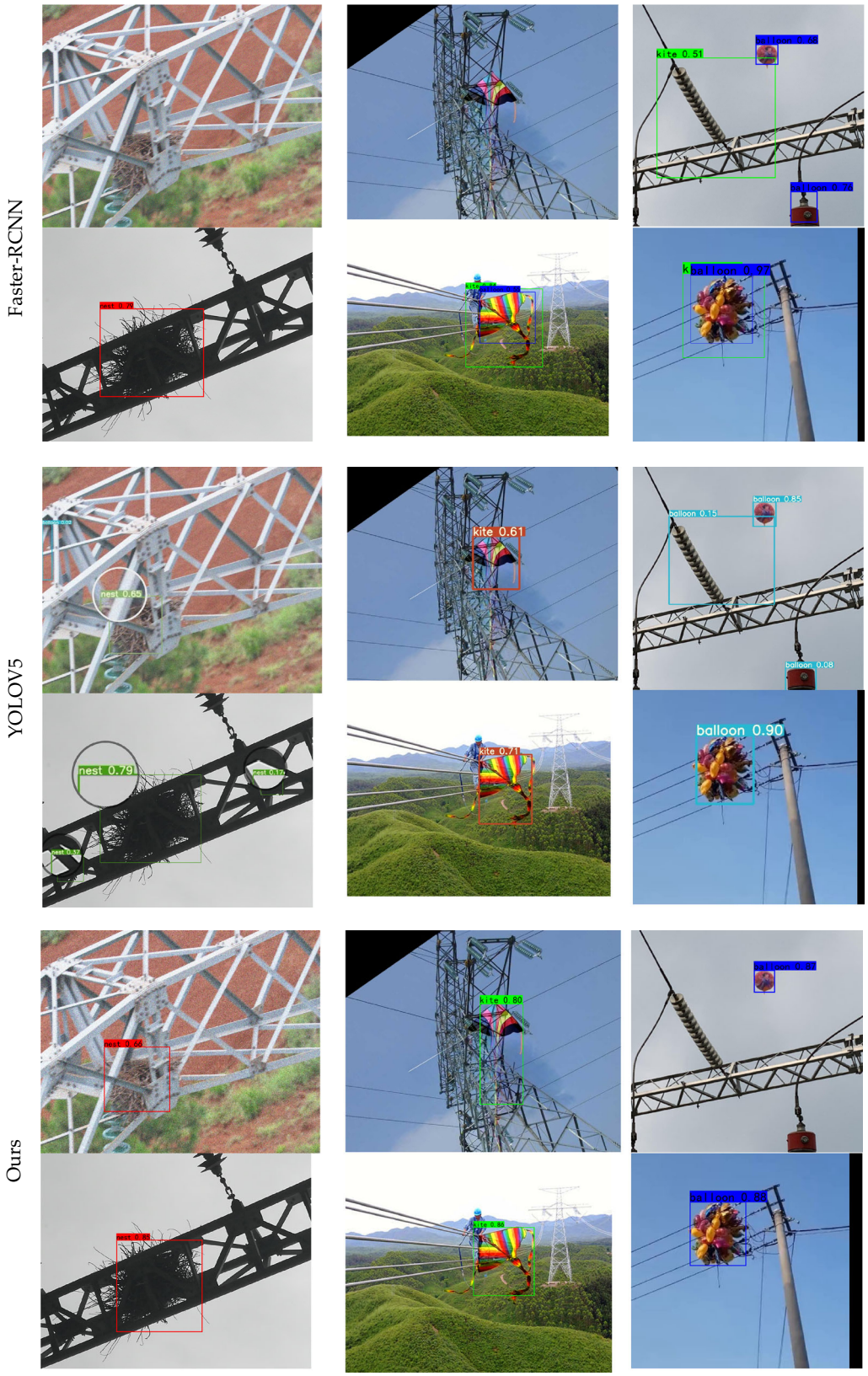


Fig. 14 Detection results of different network models. **a** The nest detection results of different network models. **b** The kite detection results of different network models. **c** The balloon detection results of different network models

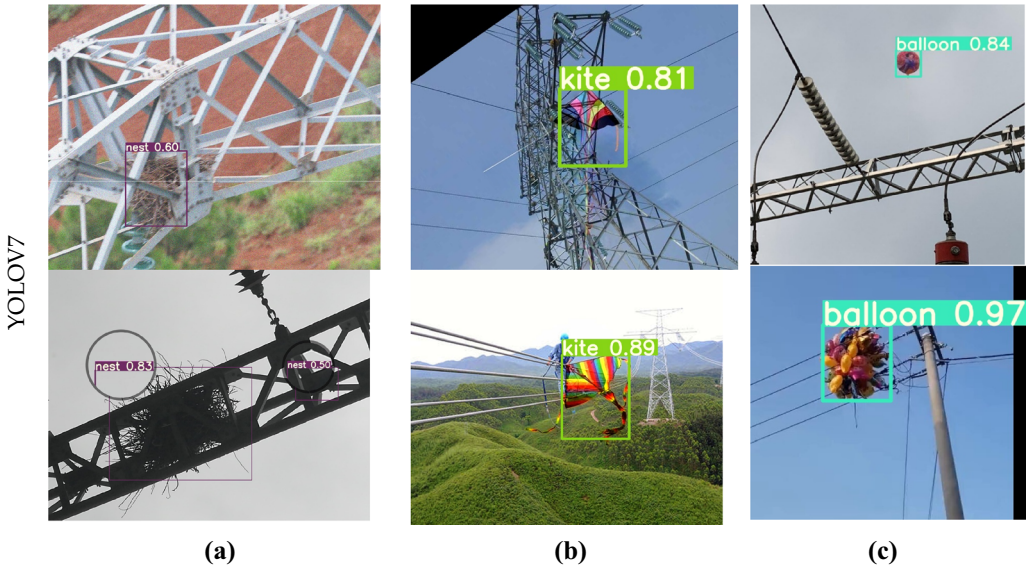


Fig. 14 continued

Author contributions CT, HD, YH, and TH contributed to methodology and conceptualization; HD and CT conceived and designed the experiments; HD, MF, and JF contributed to data curation and performed the experiments; HD analyzed the data and contributed to writing—original draft preparation; HD, CT, and YH contributed to writing—review and editing; and YH contributed to funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding This research was funded by the National Natural Science Foundation of China, Grant Number 61772033 and Anhui University Collaborative Innovation Project, Grant Number GXXT-2019-048, GXXT-2020-54.

Data availability statement The datasets analyzed during the current study are not publicly available but are available from the corresponding author on reasonable request.

Declarations

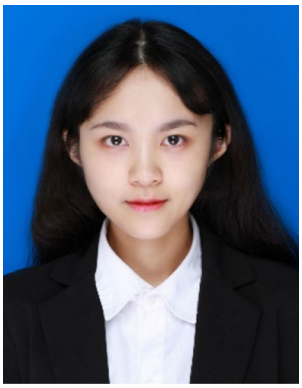
Conflict of interest The authors declare no conflict of interest. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Zhang, R., Yang, B., Xiao, W., Liang, F., Liu, Y., Wang, Z.: Automatic extraction of high-voltage power transmission objects from UAV Lidar point clouds. Remote Sens. **11**, 2600 (2019). https://doi.org/10.3390/rs11222600
2. Cheng, L., Wu, G.: Obstacles detection and depth estimation from monocular vision for inspection robot of high voltage transmission line. Cluster Comput. **22**(Suppl 2), 2611–2627 (2019). https://doi.org/10.1007/s10586-017-1356-8
3. Chen, C., Jin, A., Yang, B., Ma, R., Sun, S., Wang, Z., Zong, Z., Zhang, F.: DCPLD-Net: a diffusion coupled convolution neural network for real-time power transmission lines detection from UAV-Borne LiDAR data. Int. J. Appl. Earth Observ. Geoinf. **112**, 102960 (2022). https://doi.org/10.1016/j.jag.2022.102960
4. Nguyen, V.N., Jenssen, R., Roverso, D.: Automatic autonomous vision-based power line inspection: a review of current status and the potential role of deep learning. Int. J. Electr. Power Energy Syst. **99**, 107–120 (2018). https://doi.org/10.1016/j.ijepes.2017.12.016
5. Alhassan, A.B., Zhang, X., Shen, H., Xu, H.: Power transmission line inspection robots: a review, trends and challenges for future research. Int. J. Electr. Power Energy Syst. **118**, 105862 (2020). https://doi.org/10.1016/j.ijepes.2020.105862
6. Luo, Y., Yu, X., Yang, D., Zhou, B.: A survey of intelligent transmission line inspection based on unmanned aerial vehicle. Artif Intell Rev. **56**, 173–201 (2023). https://doi.org/10.1007/s10462-022-10189-2
7. Liu, X., Miao, X., Jiang, H., Chen, J.: Data analysis in visual power line inspection: an in-depth review of deep learning for component detection and fault diagnosis. Annu. Rev. Control. **50**, 253–277 (2020). https://doi.org/10.1016/j.arcontrol.2020.09.002
8. Chen, C., Yang, B., Song, S., Peng, X., Huang, R.: Automatic clearance anomaly detection for transmission line corridors utilizing UAV-Borne LIDAR data. Remote Sens. **10**, 613 (2018). https://doi.org/10.3390/rs10040613
9. Liu, Y., Liao, L., Wu, H., Qin, J., He, L., Yang, G., Zhang, H., Zhang, J.: Trajectory and image-based detection and identification of UAV. Vis. Comput. **37**, 1769–1780 (2021). https://doi.org/10.1007/s00371-020-01937-y
10. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, pp. 886–893 (2005). https://doi.org/10.1109/CVPR.2005.177
11. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**, 91–110 (2004). https://doi.org/10.1023/B:VISI.0000029664.99615.94
12. Cha, Y.J., You, K., Choi, W.: Vision-based detection of loosened bolts using the Hough transform and support vector machines. Autom. Constr. **71**(Part 2), 181–188 (2016). https://doi.org/10.1016/j.autcon.2016.06.008
13. Wang, J., Wang, J., Shao, J., Li, J.: Image recognition of icing thickness on power transmission lines based on a least squares Hough transform. Energies **10**, 415 (2017). https://doi.org/10.3390/en10040415

14. Duda, R.O., Hart, P.E.: Use of the Hough transformation to detect lines and curves in pictures. Commun. ACM **15**, 11–15 (1972). https://doi.org/10.1145/361237.361242
15. Hazgui, M., Ghazouani, H., Barhoumi, W.: Genetic programming-based fusion of HOG and LBP features for fully automated texture classification. Vis. Comput. **38**, 457–476 (2022). https://doi.org/10.1007/s00371-020-02028-8
16. Lu, J., Xu, X., Li, X., Li, L., Chang, C., Feng, X., Zhang, S.: Detection of bird's nest in high power lines in the vicinity of remote campus based on combination features and cascade classifier. IEEE Access **6**, 39063–39071 (2018). https://doi.org/10.1109/ACCESS.2018.2851588
17. Fan, J., Yang, X., Lu, R., Li, W., Huang, Y.: Long-term visual tracking algorithm for UAVs based on kernel correlation filtering and SURF features. Vis. Comput. **39**, 319–333 (2023). https://doi.org/10.1007/s00371-021-02331-y
18. Liu, D., Cui Y., Tan W., Chen, Y.: SG-Net: Spatial Granularity Network for One-Stage Video Instance Segmentation. arXiv preprint arXiv:2103.10284 (2021). https://doi.org/10.48550/arXiv.2103.10284
19. Wang, W., Liang, J., Liu, D.: Learning Equivariant Segmentation with Instance-Unique Querying. arXiv preprint arXiv:2210.00911 (2022). https://doi.org/10.48550/arXiv.2210.00911
20. Li, H., Dong, Y., Liu, Y., Ai, J.: Design and implementation of UAVs for bird's nest inspection on transmission lines based on deep learning. Drones **6**, 252 (2022). https://doi.org/10.3390/drones6090252
21. Zhao, W., Xu, M., Cheng, X., Zhao, Z.: An insulator in transmission lines recognition and fault detection model based on improved faster RCNN. IEEE Trans. Instrum. Meas. **70**, 1–8 (2021). https://doi.org/10.1109/TIM.2021.3112227
22. Zhang, X., Gong, Y., Qiao, C., et al.: Multiview deep learning based on tensor decomposition and its application in fault detection of overhead contact systems. Vis. Comput. **38**, 1457–1467 (2022). https://doi.org/10.1007/s00371-021-02080-y
23. Xu, L., Song, Y., Zhang, W., An, Y., Wang, Y., Ning, H.: An efficient foreign objects detection network for power substation. Image Vis. Comput. **109**, 104159 (2021). https://doi.org/10.1016/j.imavis.2021.104159
24. Sarkar, D., Gunturi, S.K.: Online health status monitoring of high voltage insulators using deep learning model. Vis. Comput. **38**, 4457–4468 (2022). https://doi.org/10.1007/s00371-021-02308-x
25. Li, H., Liu, L., Du, J., Jiang, F., Guo, F., Hu, Q., Fan, L.: An improved YOLOv3 for foreign objects detection of transmission lines. IEEE Access **10**, 45620–45628 (2022). https://doi.org/10.1109/ACCESS.2022.3170696
26. Qiu, Z., Zhu, X., Liao, C., Qu, W., Yu, Y.: A lightweight YOLOv4-EDAM model for accurate and real-time detection of foreign objects suspended on power lines. IEEE Trans. Power Deliv. (2022). https://doi.org/10.1109/TPWRD.2022.3213598
27. Cui, Y., Yan L., Cao Z., Liu D.: TF-Blender: Temporal Feature Blender for Video Object Detection. arXiv preprint arXiv:2108.05821 (2021). https://doi.org/10.48550/arXiv.2108.05821
28. Su, J., Su, Y., Zhang, Y., Yang, W., Huang, H., Wu, Q.: EpNet: Power lines foreign object detection with edge proposal network and data composition. Knowl. Based Syst. **249**, 108857 (2022). https://doi.org/10.1016/j.knosys.2022.108857
29. Wang, W., Han, C., Zhou, T., Liu, D.: Visual Recognition with Deep Nearest Centroids. arXiv preprint arXiv:2209.07383 (2023). https://doi.org/10.48550/arXiv.2209.073 YOLOv7: Trainable 83
30. Wang, C., Bochkovskiy, A., Liao, H.M.: Bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv preprint arXiv:2207.02696 (2022). https://doi.org/10.48550/arXiv.2207.02696
31. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430 (2021). https://doi.org/10.48550/arXiv.2107.08430
32. Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., Sun, J.: RepVGG: Making VGG-style ConvNets Great Again. arXiv preprint arXiv:2101.03697 (2021). https://doi.org/10.48550/arXiv.2101.03697
33. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **37**, 1904–1916 (2015). https://doi.org/10.1109/TPAMI.2015.2389824
34. Shen, X., Wang, H., Cui, T., Guo, Z., Fu, X.: Multiple information perception-based attention in YOLO for underwater object detection. Vis. Comput. (2023). https://doi.org/10.1007/s00371-023-02858-2
35. Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., Wei, F., Guo, B.: Swin Transformer V2: Scaling Up Capacity and Resolution. arXiv preprint arXiv:2111.09883 (2022). https://doi.org/10.48550/arXiv.2111.09883
36. Wang, S., Gao, Z., Liu, D.: Swin-GAN: generative adversarial network based on shifted windows transformer architecture for image generation. Vis. Comput. (2022). https://doi.org/10.1007/s00371-022-02714-9
37. Zhou, W., Wang, C., Xiao, B., Zhang, Z.: Human action recognition using weighted pooling. IET Comput. Vis. **8**, 579–587 (2014). https://doi.org/10.1049/iet-cvi.2013.0306
38. Lin, T., Goyal, P., Girshick, R.B., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, Venice, Italy, pp. 2999–3007 (2017). https://doi.org/10.1109/TPAMI.2018.2858826
39. Leslie, N. Smith.: Cyclical Learning Rates for Training Neural Networks. arXiv preprint arXiv:1506.01186 (2017). doi:https://doi.org/10.48550/arXiv.1506.01186
40. Kingma, D., Ba, J.: Adam: A Method for Stochastic Optimization. arXiv preprint arXiv:1412.6980 (2014). https://doi.org/10.48550/arXiv.1412.6980
41. Zhang, H., Hu, Z., Hao, R.: Joint information fusion and multi-scale network model for pedestrian detection. Vis. Comput. **37**, 2433–2442 (2021). https://doi.org/10.1007/s00371-020-01997-0
42. Ren, S.Q., He, K.M., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**, 1137–1149 (2017). https://doi.org/10.1109/TPAMI.2016.2577031
43. Ultralytics/yolov5. https://github.com/ultralytics/yolov5. Accessed 25 June 2020

Chaoli Tang graduated from the University of Science and Technology of China with a Doctorate in optics. Now he is a professor. He is mainly engaged in the research of atmospheric data and information technology. In the last 5 years, he has published 23 papers related to multisource Big data analysis as the first (or corresponding) author, including 18 SCI papers and 5 CSCD papers.

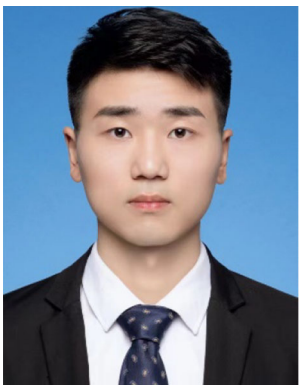


Huiyuan Dong a postgraduate student in Anhui University of Science and Technology. Her major is electrical engineering and she is mainly engaged in the research of image processing, object detection and image compression technology. In the past 5 years, she has published a paper as a collaborator in Journal of Optoelectronics Laser.

Yourui Huang secondary professor, doctoral degree, doctoral supervisor. He is currently the vice secretary of the party committee and the dean of West Anhui University, mainly engaged in the research of intelligent information processing and mine Internet of things. He has published more than 50 academic papers, more than 20 papers included by SCI and EI, and 3 academic monographs.

Tao Han a graduate student of Anhui University of Science and Technology. Now he is a senior experimentalist in Anhui University of Science and Technology, and mainly engaged in the research of object detection. In the past 5 years, he has published lots of papers as a correspondence author or collaborator.

Mingshuai Fang a graduate student of Anhui University of Science and Technology, His main research are object detection and segmentation for agricultural pest. In the past 5 years, he has published two papers as a correspondence author or collaborator. Two papers were published in Journal of Optoelectronics Laser.



Jiahao Fu a postgraduate student in Anhui University of Science and Technology. His main research directions are satellite navigation, inertial navigation, visual navigation, swarm intelligence algorithms and path planning. In the past 5 years, he has published three SCI papers as a correspondence author or collaborator. Two papers were published in journal of Agriculture and one paper was published in journal of Applied Sciences.