machines

MDPI

Article

Double-Attention YOLO: Vision Transformer Model Based on Image Processing Technology in Complex Environment of Transmission Line Connection Fittings and Rust Detection

Zhiwei Song [1], Xinbo Huang [1,2,*], Chao Ji [1] and Ye Zhang [1]

[1] School of Electronics and Information, Xi'an Polytechnic University, Xi'an 710048, China
[2] College of Mechanical and Electrical Engineering, Xidian University, Xi'an 710071, China
* Correspondence: huangxb1975@163.com

Abstract: Transmission line fittings have been exposed to complex environments for a long time. Due to the interference of haze and other environmental factors, it is often difficult for the camera to obtain high quality on-site images, and the traditional image processing technology and convolution neural networks find it difficult to effectively deal with the dense detection task of small targets with occlusion interference. Therefore, an image processing method based on an improved dark channel defogging algorithm, the fusion channel spatial attention mechanism, Vision Transformer, and the GhostNet model compression method is proposed in this paper. Based on the global receptive field of the saliency region capture and enhancement model, a small target detection network Double-attention YOLO for complex environments is constructed. The experimental results show that embedding a multi-head self-attention component into a convolutional neural network can help the model to better interpret the multi-scale global semantic information of images. In this way, the model learns more easily the distinguishable features in the image representation. Embedding an attention mechanism module can make the neural network pay more attention to the salient region of image. Dual attention fusion can balance the global and local characteristics of the model, to improve the performance of model detection.

Keywords: transmission line connection fittings; multi-scale target detection; Vision Transformer; image defogging technology; attention mechanism; model compression and optimization

check for updates

Citation: Song, Z.; Huang, X.; Ji, C.; Zhang, Y. Double-Attention YOLO: Vision Transformer Model Based on Image Processing Technology in Complex Environment of Transmission Line Connection Fittings and Rust Detection. *Machines* **2022**, *10*, 1002. https://doi.org/10.3390/machines10111002

Academic Editor: Antonios Gasteratos

Received: 19 August 2022
Accepted: 28 October 2022
Published: 31 October 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The power industry is an important pillar of the national economy. With the continuous expansion of the scale of the power grid, the number of power grid equipment units under complex environments is also increasing day by day, which requires higher requirements for the safety maintenance of power system equipment. The connecting fittings of transmission lines have the characteristics of complex structure, small parts, and difficult detection, which are exposed to complex natural environment all year round and being under the influence of haze, rain, snow, light, wind erosion and other factors, the rust phenomenon will appear. Significant rusting of metal will lead to parts falling off or failure, which will bring security risks to the stability of transmission line operation. It can be difficult for UAVs and line inspection robots to capture high-quality real-time images, which makes it difficult for feature extraction during later image processing. Therefore, it is helpful for fault diagnosis and early warning of problems in transmission lines to adopt appropriate detection methods to obtain real-time status information for connecting fittings.

Convolution neural networks are the most widely used method in the field of target detection. Target detection algorithms based on deep learning can be roughly divided into two categories. The first is the two-stage detection algorithm represented by RCNN [1] and Faster RCNN [2]. Wu et al. [3], by constructing a multi-task area recommendation network to constrain ROI, made ROI the focus of a feature extraction network. Faster

RCNN can improve its small-scale target detection ability. Mai et al. [4] used a multi-classifier fusion strategy to improve the feature learning effect of the Fast RCNN model. The second type involves the single-stage target detection algorithms represented by the SSD [5] and YOLO [6,7] series. This method regards the target detection task as a regression problem. The method is slightly lower than the two-stage detection algorithm in detection accuracy, but its real-time performance is greatly improved. Lu et al. [8], through the fusion of the SSD algorithm with two-way attention optimization, enhanced the representation of network features, and improved the detection performance of multi-scale targets. Ge et al. [9] proposed a lightweight model of the UW-YOLOv3 to solve the problem of energy consumption and storage resource limitation calculation in underwater application scenarios.

Transformer [10] is a network architecture that was first applied to machine translation tasks in the NLP domain. In addition, it has improved the ability of the traditional parallel task models, such as the RNN [11] and LSTM [12], to study the complexity of multiple tasks. Subsequent research works, such as BERT [13,14] and GPT [15], were improved on this basis, and several advanced indicators in NLP tasks were obtained. The great achievements in NLP also promote researchers to explore the possibility of applying Transformer to computer vision tasks. The research shows that Transformer plays a bridging role in the unified modeling between NLP and CV. The most significant difference between the CNN model and the Transformer model is the difference in the size of the receptive field; CNN is a generic term for conventional convolution modules. The structure based on Transformer relies on self-attention, which has advantages in capturing remote pixel information. The ViT divides the input of the image into multiple patches, each patch represents a token as the basic element of the Transformer input sequence. Although it pays more attention to global information, it ignores the role of local patterns in image space, which makes it difficult for the model to capture the spatial information within each patch. Convolutional block attention models, such as CBAM [16], SENet [17], ECA [18], and AAM [19], can effectively enhance the expression of feature objects in a complex background, and enhance the saliency of the target to be detected. This makes up for the spatial local information ignored by Transformer in the patch. The combination of the two models can significantly improve the performance of the model for multi-scale complex background target detection.

The conventional convolution operation will produce a lot of unnecessary computational redundancy, which is not desired. Model compression and optimization can alleviate the problem of large number of parameters and redundant calculation information of a complex model. Paoletti et al. [20] used the combination of GhostNet and the CNN based HSI classifier to reduce the computational cost of the model and achieved a high performance index on a small number of hyperspectral imaging classification datasets. Yue et al. [21] proposed a lightweight object detection model YOLO-GD (GhostNet and depth convolution), which reduced the inference time per image from 207.92 ms to 32.75 ms after lightweight processing. In addition, Du et al. [22] proposed a disaster prevention and safety detection model for transmission lines based on YOLOv5S, which improved the efficiency and detection accuracy of the original network by fusing BiFPN and GhostNet structures. Yan et al. [23] used the improved fast RCNN network to locate the target in the infrared image of a transmission line, which had high recognition accuracy.

This paper proposes a multi-model fusion method to improve the performance of the hybrid model. YOLOv5 was used as the baseline model for the study. In this method, the model Vision Transformer was added to the original YOLOv5 detection network to explore the development potential of multi-head self-attention in multi-scale target detection. First, Vision Transformer and GhostNet [24] were integrated into the backbone, then linear computational complexity was used to compress the model and improve its computational efficiency. At the same time, the global modeling ability of self-attention in Vision Transformer was fully utilized to increase the receptive field of the model. Although Vision Transformer reduces the input sequence information by converting images into multiple patches, it can be deployed in the CV field. However, it also exposes the disadvan-

tages of Transformer, as abandoning the traditional convolution operation will cause the model to lose the ability to capture local details, and Transformer does not have an adaptive local pattern to obtain the details of each patch. To solve this problem, we modified the neck and head parts of YOLOv5. First, the original FPN + PAN [25,26] structure was replaced by a more efficient BiFPN [27] structure, to enhance the ability of multi-scale information fusion. Then the original neck of YOLOv5 was replaced by the structure of Vision Transformer + CBAM. The addition of CBAM can effectively enhance the saliency of the target to be detected in complex environment, to improve the ability of the model to capture fine-grained information of the image. The proposed model can effectively balance the dependence between global features and local spatial information.

The main contributions of this paper are as follows:

1. A haze removal algorithm for transmission line fittings based on image processing was designed. This method enhances the resolution of the target to be detected in the original image through improved dark channel haze removal technology, which can improve the quality of the image collected outdoors and facilitate the application and deployment of the target detection algorithm.

2. A Double-attention YOLO network model was proposed as a feature extraction network. By fusing the hierarchical extraction of feature maps, it focuses on the region of interest, enhances the global receptive field of the model, improves the semantic recognition ability in complex detection tasks and reduces the feature confusion of the classification model. It also has better robustness in dealing with problems such as target occlusion, noise interference, and region offset.

3. For the first time, the structure of Vision Transformer was integrated into the convolution structure of YOLOv5, which improved the understanding of the semantic information of the graphics context of the original model to distinguish the feature representation of the model and enhance the global receptive field of the model on the feature map.

4. The attention mechanism module CBAM was introduced in the head of the model to make the neural network pay more attention to the salient details of the region to be detected.

5. The GhostNet model compression module was used to reduce the computational redundancy generated by the neural network in the feature extraction process and the model inference speed of the multi-target detection task of the transmission line connection hardware was improved through cheap linear operations.

6. Based on the proposed target detection algorithm, a condition monitoring system for transmission line fittings was developed to facilitate the application and deployment of the model.

The remaining sections of this paper are arranged as follows: In the second section, the relevant methods are analyzed theoretically. Section 2.1 is the introduction of hardware system and model deployment. Section 2.2 is the theoretical analysis of complex scene defogging algorithm based on improved dark channel prior. Section 2.3 is the introduction of the Double-attention YOLO model structure. Section 2.4 proposes an anchors clustering method based on improved K-means. The third section is the experimental part. Section 3.1 is the application of the improved dark channel prior defogging algorithm in this study. Section 3.2 is the anchors clustering method based on improved K-means and the anchors clustering experiment optimized by genetic algorithm. Section 3.3 is the performance verification experiment of the proposed Double-attention YOLO model and the comparison experiment with other advanced methods. The fourth section is the conclusion and summary.

2. The Proposed Theory

2.1. Hardware System and Model Deployment

The field deployment process of the proposed rust detection method for transmission line fittings under complex environmental conditions based on Vision Transformer and the image processing technology proposed in this paper is shown in Figure 1. This system can

collect information about the equipment status of UHV transmission lines in daytime and can resist the interference of complex external factors such as rain, snow, haze, wind, and earthquakes. First, the image acquisition device obtains the real-time status information for the connecting fittings of the transmission line, and then transmits the collected information to the ground equipment health monitoring center through the 4G communication wireless network, sorts the data into various states and the different environments into batches, and transmits them to the server for data preprocessing and enhancement. Finally, image analysis and data training and testing of the target detection task are carried out with relevant algorithms.

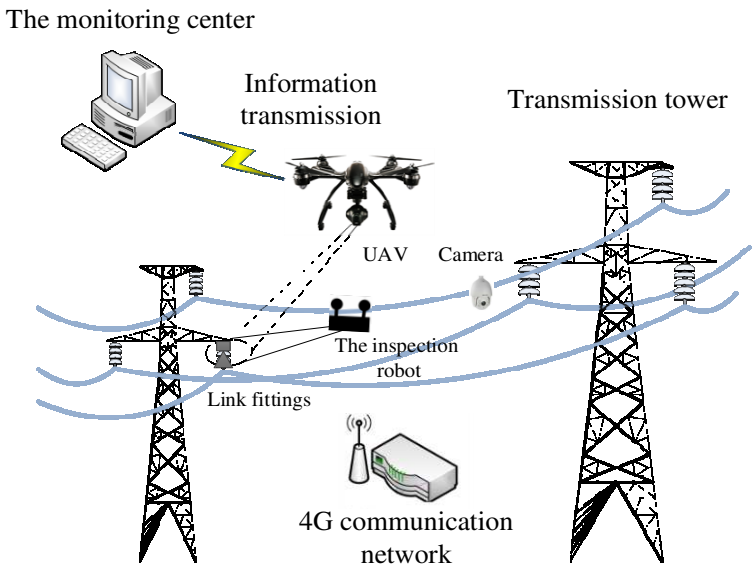


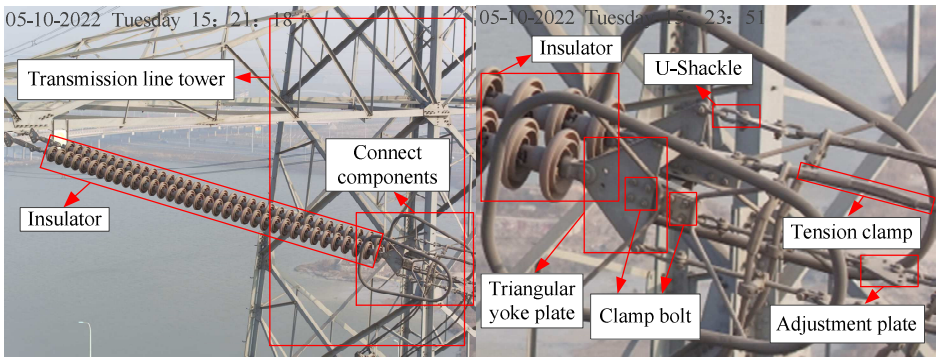
Figure 1. Block diagram of the online monitoring system.

To better adapt to various complex environmental conditions and resist external disturbance, the system selects three kinds of acquisition equipment for the deployment of the image acquisition device, as shown in Figure 2. First, is the inspection robot of the transmission line (Figure 2a), which shuttles between transmission line poles and towers with a ground wire as a running track, and high-definition cameras are installed at the front end and bottom to collect images. Second, is a high-definition camera fixed on the top of the tower (Figure 2b). This camera is an online monitoring device, which has the advantages of a wide field of vision and high capture accuracy and can transmit the status information of the transmission line connecting fittings in different environments in real time. Third, is a patrol UAV (Figure 2c), which has the advantage of high flexibility through manual control, and is suitable for fine detection of transmission line equipment with potential risks. The hardware detection equipment is connected to the ground server base station through a wireless communication module, thus forming the whole transmission line online monitoring system.

Example onsite images collected by the online monitoring system are shown in Figure 3, in which Figure 3a is a real time picture of a one-stage detection task. The targets to be detected in Figure 3b are the internal information of connecting components in Figure 3a, and there are six types of power transmission line fittings: U-Shackle (US), Triangular yoke plate (TYP), Adjustment plate (ADP), Tension clamp (TEC), Clamp bolt (CLB), Insulator (In). The figure shows all the kinds of information and position information of the two-stage detection image.

(a) The inspection robot (b) Camera (c) UAV

Figure 2. Hardware acquisition device for on-line monitoring system of transmission line connection fittings.



(a) On-site aerial images (b) Connect components

Figure 3. Images captured on site.

In addition, we integrate the overall algorithm design ideas and draw the overall algorithm structure diagram as shown in Figure 4.

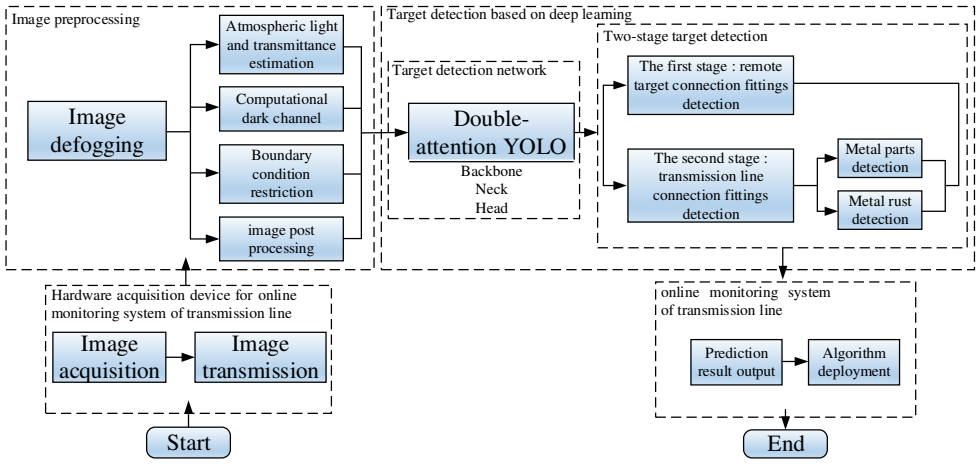


Figure 4. Overall deployment structure diagram of the algorithm.

Inspection of Figure 4 shows that the overall design of the algorithm is divided into four parts, namely image acquisition and transmission, image defogging, connected hardware target detection, online monitoring of the deployment of the system algorithm. The image acquisition and transmission module are mainly completed by the system hardware. The image defogging algorithm is divided into a defogging stage and an image post-processing stage. The image post-processing stage improves the defogging effect through a series of segmentation and edge detection methods. The two-stage object detection algorithm is based on our proposed Double-attention YOLO model, which is

the core module in the four stages of the whole algorithm. The last stage will output the test results to the transmission line online monitoring system for unified deployment and early warning. This completes the overall design of the whole rust detection method for connection fittings or parts on transmission lines.

2.2. Fog Removal Algorithm for Complex Scenes Based on Improved Dark Channel Prior

Due to the influence of fog and haze, it can be difficult to detect a fault in transmission line fittings by UAV or inspection robot. Therefore, it is necessary to correct the image by designing an auxiliary defogging method that is suitable for these complex scenes before the implementation of target detection algorithm.

In this paper, based on the improved dark channel prior criterion defogging algorithm, the images of outdoor transmission lines collected by the equipment on foggy days were defogged. The flow chart of the defogging algorithm is shown in Figure 5, where Figure 5a shows the estimation of atmospheric light and transmittance, Figure 5b calculates the dark channel for the down sampling process and Figure 5c is the boundary condition restriction based on the radiation cube criterion. Figure 5 shows the overall flow chart of defogging algorithm.

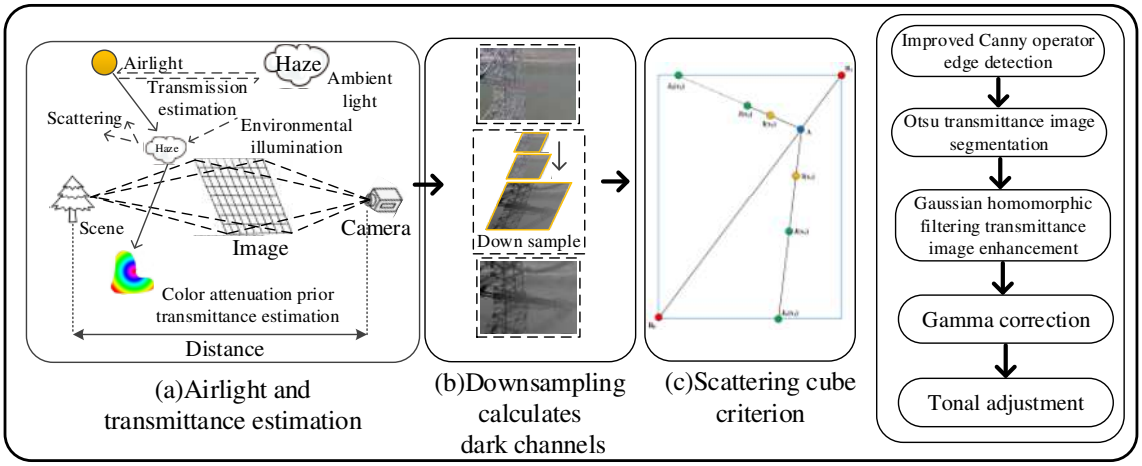


Figure 5. Overall flow chart of defogging algorithm.

The original dark channel defogging algorithm has obvious color distortion in the recovery effect of the sky area. Therefore, we used the color attenuation prior method to establish the scene depth model to eliminate this phenomenon, and estimated the transmittance based on the scene depth model. The specific implementation process shown in Figure 5 is as follows: first, the atmospheric physical model of light scattering by suspended particles in the air is obtained, and the maximum value of the pixel set in each channel of the RGB image is found to estimate atmospheric light. The transmittance is obtained by the color attenuation prior method [28] and scene depth estimation. According to the conversion relationship between RGB and HSV, the color space model is shown in the Formula (1):

$$S(x) = \frac{G(x)}{\max\limits_{c\in\{r,g,b\}} I^c(x)}, \; G(x) = \max_{c\in\{r,g,b\}} I^c(x) - \min_{c\in\{r,g,b\}} I^c(x) \tag{1}$$

where $S(x)$ is the saturation channel. If $D(x)$ represents the depth of the scene, then $G(x)$ is negatively correlated with it $d(x) = 1 - G(x)$, I is the three dimensional color vector in RGB, $c \in \{r, g, b\}$ is the color channel index, and the expression of transmittance t' is shown in Formula (2):

$$t' = e^{-\eta d(x)} = e^{-\eta[1-G(x)]} \tag{2}$$

where η is the estimate parameters, it varies with the selection of images. The Formula (3) is given as follows:

$$t' = 1 - \omega \min_{y \in \Omega(x)} (\min_{c \in (r,g,b)} \frac{I^c(y)}{A^c}) \tag{3}$$

where A is ambient light, ω is the image removal rate (ω = 0.95). The original dark channel prior knowledge algorithm calculates the dark channel value for each pixel, which increases the computational time cost of large-scale images. We find that the dark channel image has a low spatial frequency, so we can use the image down sampling method (using 1×1 patch instead of 15×15 patch) to calculate the dark channel, and then use interpolation to restore it to its original size. Therefore, Formula (3) can be refined into Formula (5) by Formula (4):

$$T_J(x) = 1 - \lambda \frac{(\min_{c \in \{r,g,b\}} (\frac{I^c(x)}{A^c}) - \min_{x \in \Omega} (\min_{c \in \{r,g,b\}} (\frac{I^c(x)}{A^c})))}{(\max_{c \in \Omega} (\min_{c \in \{r,g,b\}} (\frac{I^c(x)}{A^c}) - \min_{x \in \Omega} (\min_{c \in \{r,g,b\}} (\frac{I^c(x)}{A^c})))} \tag{4}$$

$$t' = (1 - \omega \min_{c \in \{r,g,b\}} (\frac{I^c(x)}{A^c}))/T_J(x) \tag{5}$$

where λ is the adjustment parameters of light T_J, Ω is the entire image, $1 - \lambda \leq T_J(x) \leq 1$ (λ = 0.5). The original defogging algorithm has limited ability to restore edge detail information and does not smooth the edge information of the restored image. This results in the indiscriminate mixing of the restored image edge and non-edge information, thus affecting the image restoration effect. In the process of transmission calculation, some edge information will be lost when using down sampling. Therefore, the image boundary is restricted and smoothed by the boundary constraint and Gaussian homomorphic filtering. The boundary restriction based on the radiation cube criterion can reduce the complexity of dark channel calculation and avoid the occurrence of ladder phenomenon $B_0 \leq J(x) \leq B_1$. Through the atmospheric scattering model, we can get the boundary condition of transmittance image as follows $B_0 \leq boundary \leq t' \leq B_1$, boundary is the boundary constraint of the transmittance image. The solution of the boundary constraint is shown in Formula (6):

$$Boundary = \min\{ \max_{c \in \{r,g,b\}} (\frac{A_c - I_c(x)}{A_c - B_0^c(x)}, \frac{A_c - I_c(x)}{A_c - B_1^c(x)})\} \tag{6}$$

According to Formula (6), B_0 and B_1 are the upper and lower boundaries of the scene image, respectively. Constraints on $I(x)$ observation intensity and $J(x)$ scene brightness in a certain range by boundary. When the scene scattering problem is alleviated, the equivalent dark channel image with the ladder effect removed is obtained. After that, the boundary condition of the image is determined by using the improved region detection operator and the region detection rate. Then, the edge information of the transmittance image is enhanced by Gaussian homomorphic filtering, and the internal information is smoothed. The expression is shown in Formula (7):

$$H(u,v) = (a_H - a_L)[1 - e^{(-c(\frac{D(u,v)}{D_0})^{2n})}] + a_L \tag{7}$$

where $H(u,v)$ is the transfer function of Gaussian filter, a_H, a_L is the gain coefficient of high and low frequency, c is the sharpening constant, $D(u,v)$ represents the distance of frequency (u,v) to the filter center, D_0 is the cutoff frequency, the edge region at the abrupt change of depth of field is obviously enhanced by Gaussian homomorphic filtering.

To solve the halo phenomenon in the area with higher brightness after dark channel processing, the γ function is used to modify the envelope curve of the gray histogram of the transmission image. The correction function is shown in Formula (8):

$$h = t \times f^{\gamma} \tag{8}$$

where t is the amplitude proportional coefficient, c = 1 in this paper, f is the histogram curve to be fitted, γ is the correction index, this paper takes γ = 0.5, 1.5, 2.5. Finally, the haze free image is restored by atmospheric scattering model $J^c(x)$ and the dark image is adjusted by the tone adjustment function (9):

$$Jp^c(x) = \frac{0.01 L_{d\max}}{\lg(J^c_{\max}(x)+1)} \times \frac{\ln(J_c(x)+1)}{\ln(2+8[\frac{J^c(x)}{J^c_{\max}(x)}]^{\frac{\ln a}{\ln 0.5}})}, \; c \in \{r,g,b\} \tag{9}$$

where $Jp^c(x)$ is the output function after tone adjustment, $L_{d\max}$ is the maximum brightness value, A is the bias parameter, which is used to adjust the details of the dark area, $J^c_{\max}(x)$ is the maximum pixel value of $J^c(x)$.

2.3. Model Structure of Double-Attention YOLO

The overall network structure of the Double-attention YOLO model is composed of a backbone and a head. Before the feature extraction of the backbone part, the data enhancement operation is first performed by Maxup [29]. As shown in Figure 6, the backbone part of the model is composed of the GhostNet bottleneck and a convolution structure and then outputs to the Vision Transformer block and the SPPF layer. The function of the GhostNet module is to reduce the redundancy of model calculation information through linear operations to achieve model compression; the detailed structure will be described in Section 2.3.3. The function of Vision Transformer block is to enhance the global receptive field of the model on the feature map, capture more richer and stronger semantic information for subsequent input, and its detailed structure and working principle will be introduced in Section 2.3.1.

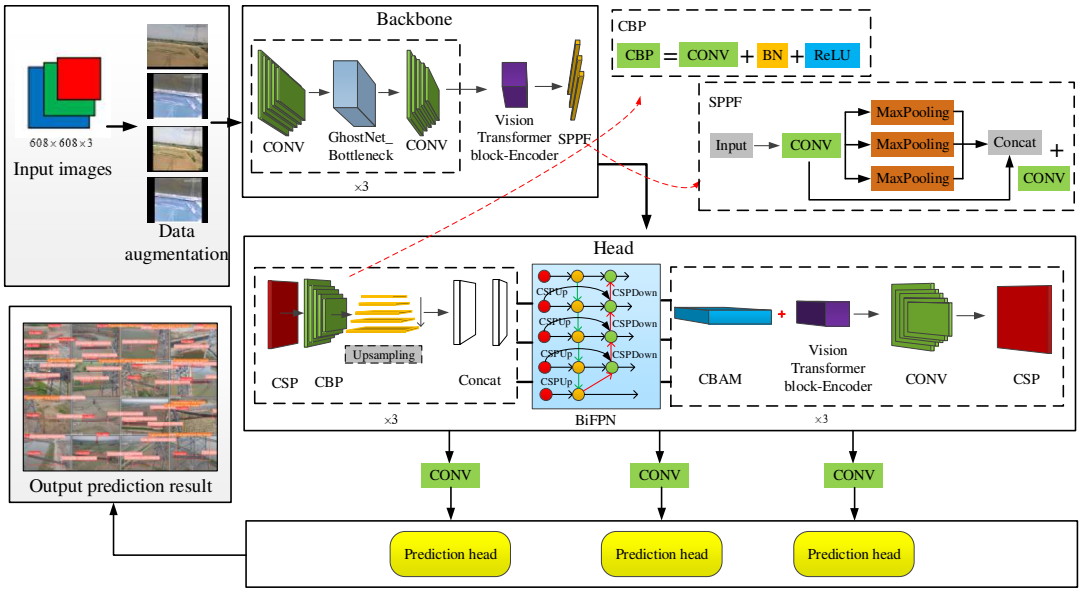


Figure 6. The structure diagram of Double-attention YOLO.

The SPPF is a spatial pyramid pooling layer, which improves the multi-scale information fusion ability of the model by transforming input features of different sizes into vector information of specific dimensions. The specific structure of the SPPF is given on the right side of Figure 6. The main structure of the SPPF is convolution and concat operation of max pooling. The head's network structure abandons the FPN + PAN structure of the original YOLOv5 and uses the convolution output feature map down sampling to enhance the receptive field. The CBP structure is CONV + BN + ReLU, and then through the BiFPN feature, fusion structure, the input information of different sizes is fully extracted for fusion operation. The model structure and working principle of the BiFPN will be introduced

in Section 2.3.3. Finally, the multi-scale target recognition ability of the model will be enhanced through the CBAM channel spatial attention mechanism module and the Vision Transformer block. The working principle of the CBAM attention mechanism unit will be introduced in Section 2.3.2. In the prediction end of the model, we use $(1 - IOU)$ and the improved K-Means algorithm to get nine kinds of anchors of different sizes by clustering, in which every three anchors of similar size are classified into a group, and the prediction of large, medium, and small size are, respectively, predicted, the improved K-Means anchor clustering process will be introduced in Section 2.4.

2.3.1. The Architecture of Vision Transformer

Compared with CNN, which is widely used in the CV field, Vision Transformer has the advantage that there is no sign of saturation as the depth of model deepens and the size of datasets increase. It can process any length of sequence information within the scope of memory. In small and medium sized image recognition tasks, convolutional neural networks, represented by ResNet are still the most advanced. The spatial locality and two dimensional neighborhood features run through all network sublayers, while only the MLP layer in ViT [30] has the above characteristics, and the role of self-attention on the feature graph is global, which results in Transformer lacking some relevant inductive bias compared with convolutional neural network. As the amount of data increases, the advantages of Transformer begin to show gradually. This paper uses Vision Transformer to replace part of the network structure of YOLOv5 to obtain the best recognition performance. The network structure diagram of Vision Transformer is shown in Figure 7.

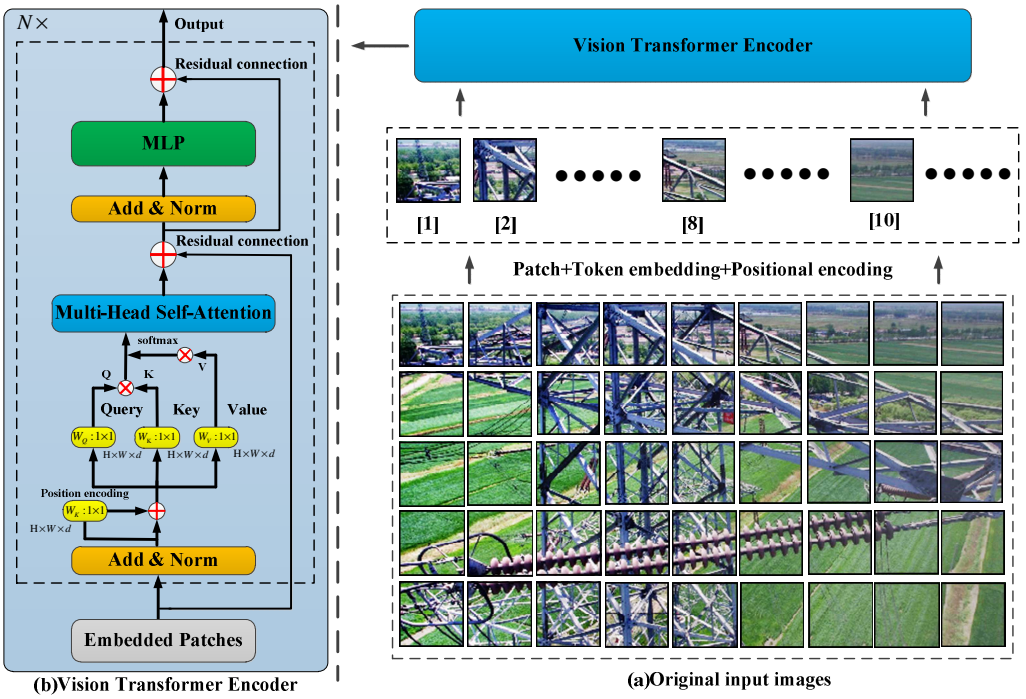


Figure 7. Vision Transformer network structure diagram.

Figure 7 shows that Vision Transformer divides the input image into a certain number of patches (Figure 7a) and transmits them to token embedding and positive encoding as sequence information. The pure encoder structure based on Transformer is the 1D vector embedded in the sequence input, which solves the problem that it is difficult to obtain the global characteristics of the existing neural network. The image position information can be effectively preserved by $1D$ position embedding, which can be learned by linear layer n. When global self-attention is executed in entity objects, the $O(n^2d)$ computational complexity of Transformer is significant. The encoder of Vision Transformer is composed of two independent sub-layers: the MLP layer and the multi-head self-attention (MSA)

network. On every floor l, the following operations are performed: input $I^{l-1} \in R^{L \times C}$ the resulting (Q, K, V) triples are used as the input of self-attention. Among them:

$$Q = I^{l-1},\ K = I^{l-1}W_K,\ V = I^{l-1}W_V \tag{10}$$

where, W_Q, W_K, W_V are the weights of three related linear mapping vectors, corresponding to the yellow part of Figure 7b, d is the dimension of feature vector. The process of self-attention (SA) can be expressed as follows:

$$SA\ (I^{l-1}) + soft\max(\frac{I^{l-1}W_Q(ZW_K)^T}{\sqrt{d}})\ (Z^{l-1}W_V) \tag{11}$$

where n self-attention modules are connected in series to form a multi-head mechanism and are used as output $MSA(I^{l-1}) = [SA_1(I_{l-1}); SA_2(I_{l-1}); \ldots; SA_n(I_{l-1})]W_O$, where $W_O \in \mathrm{R}^{md} \times C$, and the value of d is C/m. The output of MSA module and residual connection structure are used as the input of MLP layer.

2.3.2. Attention Mechanism Unit

The multi-head self-attention (MSA) network can search the whole region of the feature map by global self-attention and enhance the receptive field of the model. To enhance the saliency of the target to be detected in a complex background environment, this paper uses the channel and spatial attention mechanism CBAM and MSA to realize the overall perception of image size features, we add the CBAM module to the backbone and the neck of YOLOv5 network structure to verify the importance of the attention mechanism in different positions of the model. The working principle of the CBAM module is in a certain input characteristic graph $F = \mathbb{R}^{\mathrm{C} \times W \times \mathrm{H}}$. We infer the attention map along the 2D dimension (channel and space) and optimize each other with its corresponding feature map. The calculation formula for the channel information attention mechanism is shown in Formula (12):

$$M_{\mathrm{C}}(F) = \sigma(W_1(W_0(F^c_{avg})) + W_1(W_0(F^c_{Max}))) \tag{12}$$

where σ is the sigmoid nonlinear activation function, $W_0 \in \mathbb{R}^{c/r \times c}$ and $W_1 \in \mathbb{R}^{c \times c/r}$, W_0 and W_1 represent the hidden weight and output weight in MLP layer, respectively, and their input uses shared weight W_0 and W_1, F^c_{avg} and F^c_{Max} mean that using average pooling and max pooling to generate feature maps on the corresponding space, r represents reduction rate. To compensate for the deviation of channel attention on location information, the spatial attention module is introduced, and its calculation formula is shown in Formula (13):

$$M_S(F) = \sigma(f^{7\times7}([F^s_{avg}; F^s_{Max}])) \tag{13}$$

where $f^{7\times7}$ is a 7×7 convolution operation, F^s_{avg} and F^s_{Max} represent the global average pooling feature and the maximum pooling feature on the channel, respectively.

2.3.3. The Principle of the YOLOv5 Algorithm and BiFPN Feature Fusion

The YOLOv5 model adds the Focus module and CSPNet (cross stage partial network) [31] structure based on the YOLOv4 model. Its input sample is 640×640. In the pre-processing process, the samples are filled adaptively. Only simple splicing of input features may lead to mismatching of feature mapping, and some semantics which are helpful for context information modeling may be lost, which will lead to the degradation of model performance. Therefore, it is very important to introduce a set of learnable weight factors to characterize the difference of channel information, the top-up and bottom-down bidirectional multi-scale feature fusion is repeated to enhance the communication between the strong spatial information and semantic information of the model. The high-resolution feature map can obtain the feature mapping with fuzzy spatial information but complete semantic information through the up-sampling operation, and then the bidirectional path

feature mapping with the same spatial level is connected through the horizontal connection. The unidirectional transmission of feature information will restrict the traditional FPN structure and cause the network to ignore some important information. Based on this, we propose a new model called PANet, which adds a bottom-up path aggregation network to the FPN to enhance the cross-scale fusion ability of the model. In the BiFPN structure, an extra edge is added between the input node and the output node to ensure the cross-scale computation without additional parameters and computation. At the same time, to avoid the performance degradation caused by simple feature splicing, BiFPN introduces a set of learnable weight factors to represent the weight proportion of different input features. The PANet aggregates multi-scale feature information. The aggregation calculation process of the N-level model is shown in Formula (14):

$$\begin{aligned} {P_n}^{td} &= Conv(P_n^{in} + \mathrm{Re}size(P_{n+1}^{td})) \\ {P_n}^{out} &= Conv(P_n^{td} + \mathrm{Re}size(P_{n-1}^{out})) \end{aligned} \tag{14}$$

where Re*size* is the sampling operation matched with the corresponding resolution,*Conv* is convolution operation for processing features, P_n represents the feature level of layer n at a specific resolution. The n-th aggregation calculation process of BiFPN represented by learnable weight and bidirectional cross scale normalized feature fusion is shown in Formula (15):

$$\begin{aligned} P_n^{td} &= Conv(\tfrac{w_1 \cdot P_n^{in} + w_2 \cdot \mathrm{Re}size(P_{n+1}^{in})}{w_1 + w_2 + \gamma}) \\ P_n^{out} &= Conv(\tfrac{w'_1 \cdot P_n^{in} + w'_2 \cdot \mathrm{Re}size(P_n^{td}) + w'_3 \cdot \mathrm{Re}size(P_{n-1}^{out})}{w_1 + w_2 + w_3 + \gamma}) \end{aligned} \tag{15}$$

where Each channel layer P_n is multiplied by the corresponding weight factor ω_i to obtain the optimal feature scale representation. Then divide the weighted sum of each scale information to fit the original input, γ is the adjustment factor, the feature map is mapped to the sampling path by convolution.

2.3.4. Ghost Net Model Compression

In the target detection task of power transmission line fittings, due to the complexity of the color range in the discrimination information in the image, many repeated feature maps will be generated. The model calculation redundancy generated by these data will increase the consumption of energy and hardware resources. In terms of model parameters, the redundant computation information is generated by model convolution. In this paper, we used Transformer instead of the partial convolution structure and introduced a lightweight module GhostNet to remove unnecessary model parameters and computation in the backbone network Double-attention YOLO. The convolution process diagram of the Ghost module is shown in Figure 8a, and the schematic diagram of the Ghost bottleneck with different steps is shown in Figure 8b, which is like the residual connection in the ResNet [32] network.

Further inspection of Figure 8 shows that the GhostNet module is composed of a Ghost convolution stack. It generates a small number of feature graphs from conventional convolution through a small filter and then generates a new feature map like it by a series of linear operations. The combination of the two groups of feature graphs is the result of the model output. If the size of the input feature map is $H \times W \times c$, and the corresponding output size is $H' \times W' \times n$, the size of convolution kernel is $k \times k$. By convolution $P_{H' \times W' \times m}$ produce $H \times W \times c \times m \times H' \times W'$ and then through linear transformation Φ_i, the ghost characteristic graph is generated, and the calculation formula is shown in Formula (16):

$$n = m \cdot s \; ; m(s-1) = \frac{n}{s(s-1)} \tag{16}$$

where m is the number of channels in the feature graph, s is the number of linear transformations, and n is the number of generated feature graphs,$(s-1)$ is the number of effective linear transformations under identity mapping. After the original feature map

and ghost feature map are spliced, the output result will be displayed. When $n >> m$, the Ghost convolution greatly reduces the computational complexity of the model compared to the original conventional convolution through cheap linear operation, thus significantly improving the model reasoning speed of the multi-objective detection task of transmission line connection fittings.

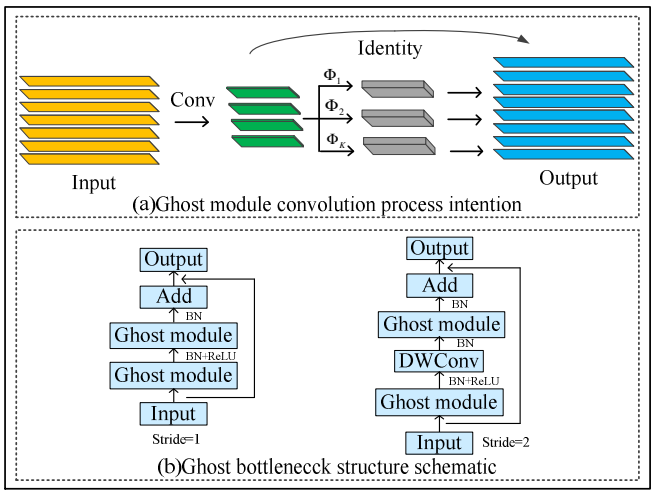


Figure 8. Ghost Net structure diagram.

2.4. Improved K-Means Anchors Clustering

The initial anchor frame of the YOLOv5 algorithm is based on the default value generated by K-Means clustering or automatically set according to the recall rate. The prediction box generated by the system adaptively cannot match the diversity of the targets in the datasets, which leads to the fluctuation of the output coordinate prediction, so the positioning accuracy will also be reduced. The Vision Transformer–YOLOv5(ViT-YOLOv5) model proposed in this paper firstly imposes constraints on the prediction boundary through anchors, and the coordinate transformation relationship of the bounding box is shown in Figure 9.

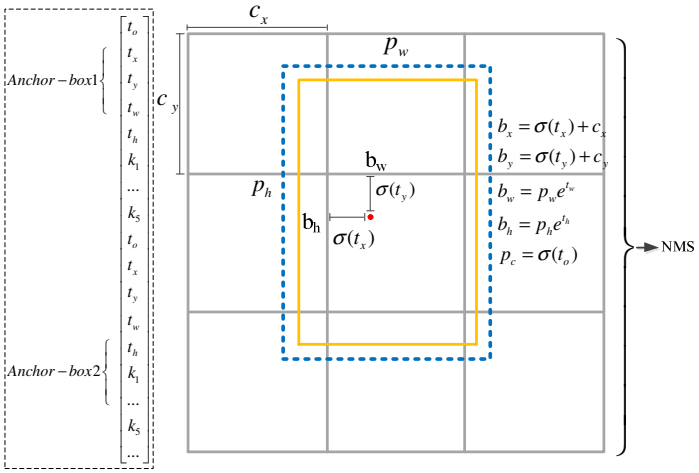


Figure 9. Bounding box coordinate transformation relation.

Among them, $\sigma(\ )$ is the sigmoid operation, by compression t_x, t_y reach $[0, 1]$ interval to enhance the positioning accuracy of the bounding box, t_w, t_h is the prediction scale, t_o is confidence, c_x, c_y is the grid coordinates of the feature map, p_c is the scaling factor, p_w, p_h are the width and height for anchors, t_x, t_y, t_w, t_h can be regarded as the learning prediction target. The center of the boundary box is placed in the grid of the second row and the second column. The prediction boundary box with low score is filtered out by

confidence threshold, and the prediction box is obtained by non-maximum suppression (NMS). To reduce the bias caused by anchors clustering, we used the $(1 - IOU)$ instead of Euclidean distance in the original algorithm, the new distance measure function is shown in Formula (17):

$$d(boxes, anchors) = 1 - IOU(boxes, anchors) \quad (17)$$

where *boxes* are labeled boxes, and *anchors* are formed by cluster centers, $IOU(boxes, anchers)$ is the intersection and merged ratio of annotation boxes and cluster centers, and the larger the *IOU* between bounding box and anchor, the closer the metric distance is. Taking *IOU* as the measurement standard, a more suitable boundary box can be found through nine anchors. The anchors calculated by the K-Means clustering algorithm are then randomly mutated by the genetic algorithm [33], the anchor results are optimized by 1000 iterations to better adapt to the target scale.

3. Experimental Results

3.1. The Application of the Improved Dark Channel Prior Defogging Algorithm in this Study

The software environment of the experiment is Matlab R2016a, and the hardware environment is Intel (R) Core(TM)i7-11700F@2.50 GHz 16.0 GB RAM. The experimental details of the improved dark channel prior algorithm were as follows: the original fog image is shown in Figure 10a, and its dark channel image in Figure 10b. The dark channel with fog image has higher intensity in the heavy haze area, and the dark channel area with higher intensity can be equivalent to the thick haze area. The transmittance map of foggy images was obtained through the dark channel prior theory and atmospheric physics model, as shown in Figure 10c. The Canny operator [34] is an effective edge detection algorithm. Its function is to eliminate some irrelevant interference and retain the target edge information by filtering noise reduction, differential calculation amplitude, non-maximum suppression, and lag threshold. In this research, the Canny operator was used to detect the edge of the transmission image, and the edge information of the transmission image was extracted to separate the edge region and the non-edge region. The edge detection result of the Canny operator is shown in Figure 10d. The analysis of the test results shows that the detected edge information was infinitely close to the target edge and satisfies the single pixel edge condition, and the edge information of insulators and connecting fittings in the image was effectively highlighted.

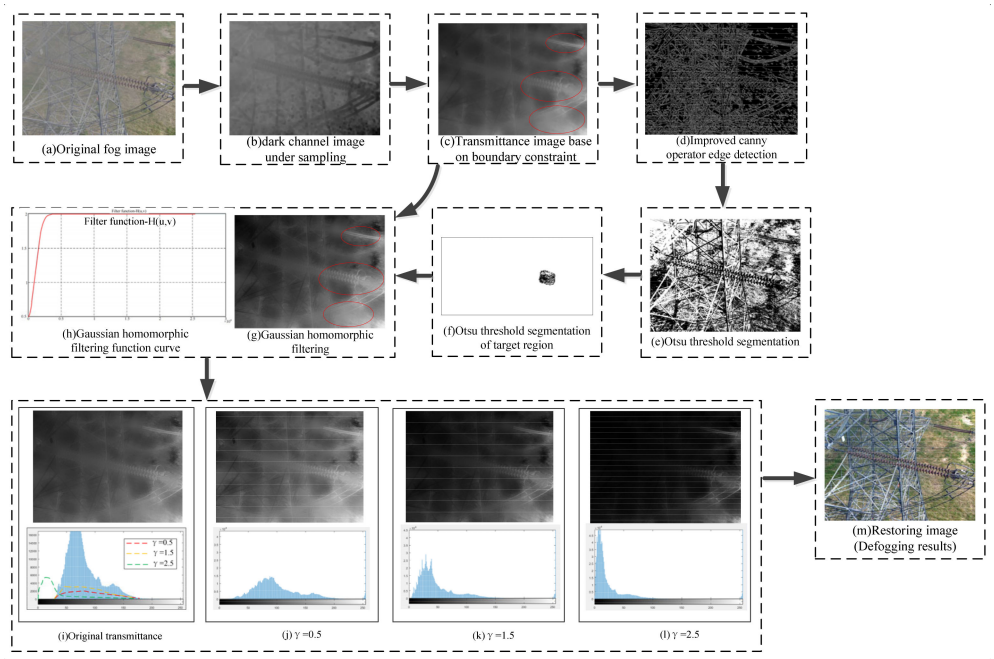


Figure 10. Experimental results of improved dark channel defogging algorithms.

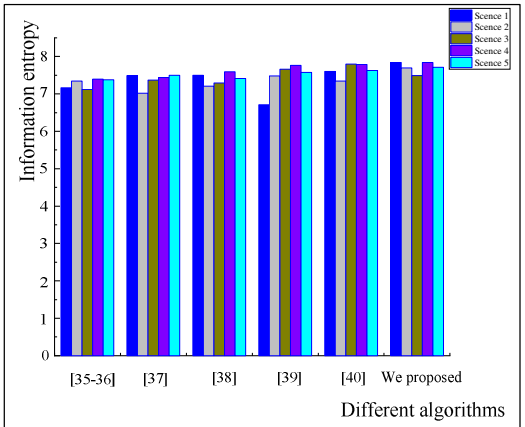
Otsu [35] is an algorithm for determining the threshold of image binarization segmentation. This method is not affected by the brightness and contrast of the image, because variance is a measure of the uniformity of gray distribution. The larger the inter-class variance between the background and the foreground, the greater the difference between the two parts of the image. When some foreground is misclassified as background or some background is misclassified as foreground, the difference between the two parts will be smaller. Therefore, the segmentation that maximizes the inter-class variance means the minimum misclassification probability. The Otsu method was used to segment the non-edge region of the target part to realize the positioning compensation of the highlighted area of the target. The segmentation results (Figure 10e) show that the whole transmittance map was traversed through the segmented binary image. It can realize the next step correction only in the specific target high background area (insulator string connection fittings), and other positions are not processed. Figure 10f is the threshold result of the binary segmentation of the connection fittings image.

The edge information of the transmittance image was enhanced by Gaussian homomorphic filtering. The processing results are shown in Figure 10g. The position of the red circle in Figure 10c,g shows that the smooth edge operation makes the image as close to the original image as possible and avoids the halo phenomenon on the basis of maintaining edge information. The edge details of the insulator and its connecting fittings in Figure 10c are kept intact and clear. The filtering result curve is shown in Figure 10h—the high frequency part of the frequency domain curve was strengthened and the low frequency part was weakened. To deal with the problem of edge information blur caused by uneven illumination and achieve accurate compensation for the target area and improve the anti-noise performance γ = 0.5, 1.5, 2.5 were used as correction coefficients to modify the envelope of the fitted transmission gray histogram. The original transmittance and gamma corrected gray image and their histogram are shown in Figure 10i–l. The modified gray envelope curve is reflected in the histogram, as shown by the color line in Figure 10i. Analysis shows that gamma correction can reduce the gray value of the area with larger gray value in the transmittance image, to realize the gray correction of specific area. The γ values need to be selected by experience based on different fog types; we found the value γ = 1.5 was the best. Finally, the corrected transmittance image was processed by hue adjustment function to obtain the restored image, as shown in Figure 10m. An ideal defogging effect can be achieved in terms of color and saturation.

To verify whether the restoration effect of the defogging algorithm in this paper was universal, six kinds of current advanced defogging algorithms were selected for comparative experiments in five different scene environments. Four kinds of image quality evaluation indexes: information entropy, Peak Signal to Noise Ratio (PSNR), MSE and average gradient were used to compare the implementation effects of various algorithms in different environments. The statistical results are shown in Figure 11a–d, where [35,36] was the Dark channel prior algorithm, Ref. [37] was the Multiscale Retinex algorithm, Ref. [38] was the Fuzzy method algorithm, Ref. [39] was the Color attenuation prior algorithm, Ref. [40] was the Adaptive histogram equalization algorithm, and the last column was our proposed method. The dominant data at the bottom of the histogram is displayed in bold font and the visualization results of the various algorithms are shown in Figure 12.

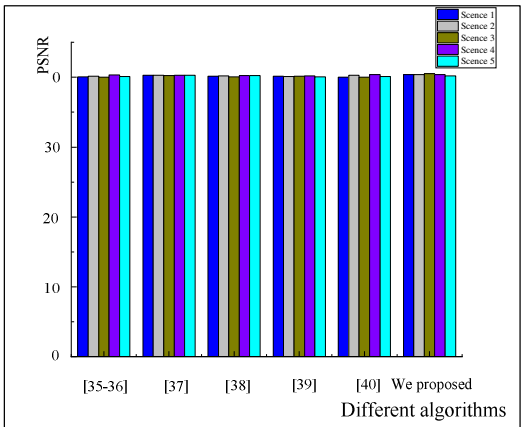
The comparison results of information entropy shown in Figure 11a reveal that the method proposed in this paper is effective in measuring the average information rate of images in most scenarios. However, the information entropy of the adaptive histogram equalization method in Scene 3 reached 7.7998, which was better than the method in this paper, because the brightness of the grassland background in the scene made the estimation of transmittance biased. The results show that our algorithm was still superior to the other algorithms in reducing the amount of noise in small scenes. The MSE in Figure 11c counts the pixel differences between different categories of images; the root mean square error of this method was the lowest in five scenarios. The average gradient of the image was analyzed by the change rate of the gray value. Figure 11d shows that the Multiscale Retinex

method had a significant defogging effect on fogging in some scenes, and the gradient value of image restoration was equivalent to that of the method in this paper. The average gradient value of the two methods in Scene 1 reached 1.862 and 1.898, respectively, which was much higher than the other four methods, which indicates that this method was good at eliminating the influence of water background on the restoration image halo.



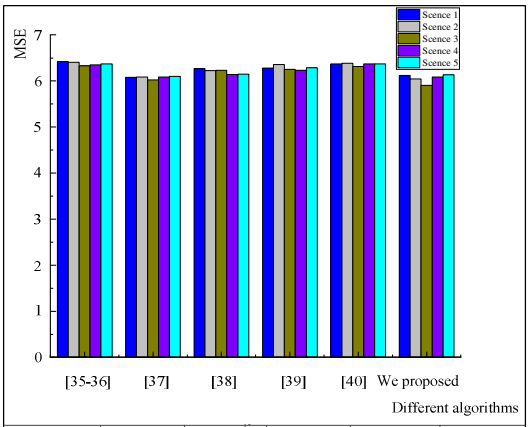
	Scence 1	Scence 2	Scence 3	Scence 4	Scence 5
[35-36]	7.1667	7.3453	7.1202	7.3956	7.3795
[37]	7.4863	7.0193	7.3689	7.4378	7.4951
[38]	7.4990	7.2119	7.2937	7.5940	7.4130
[39]	6.7087	7.4821	7.6651	7.7644	7.5746
[40]	7.6046	7.3401	7.7998	7.7897	7.6261
We proposed	**7.8375**	**7.6900**	7.4928	**7.8428**	**7.7080**

(a) Information entropy comparison results of different algorithms.



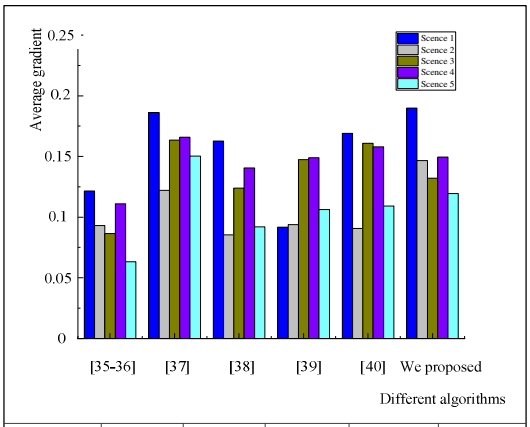
	Scence 1	Scence 2	Scence 3	Scence 4	Scence 5
[35-36]	40.0628	40.1647	40.0214	40.3310	40.0893
[37]	40.2901	40.2877	40.2603	40.2801	40.2792
[38]	40.1653	40.1896	40.0838	40.2563	40.2413
[39]	40.1565	40.0966	40.1692	40.1862	40.0472
[40]	40.0398	40.2831	40.0191	40.3908	40.0903
We proposed	**40.4130**	**40.4096**	**40.5281**	40.3903	40.2212

(b) PSNR comparison results of different algorithms.



	Scence 1	Scence 2	Scence 3	Scence 4	Scence 5
[35-36]	6.4203	6.4064	6.3310	6.3499	6.3702
[37]	6.0818	6.0857	6.0242	6.0831	6.0976
[38]	6.2736	6.2247	6.2331	6.1391	6.1506
[39]	6.2855	6.3594	6.2543	6.2297	6.2858
[40]	6.3720	6.3821	6.3121	6.3679	6.3687
We proposed	**6.1291**	**6.0434**	**5.9057**	6.0834	**6.1335**

(c) MSE comparison results of different algorithms.



	Scence 1	Scence 2	Scence 3	Scence 4	Scence 5
[35-36]	0.1216	0.0931	0.0864	0.1109	0.0632
[37]	0.1862	0.1220	0.1635	0.1658	0.1502
[38]	0.1627	0.0854	0.1239	0.1405	0.0926
[39]	0.0918	0.0939	0.1474	0.1489	0.1062
[40]	0.1691	0.0906	0.1609	0.1579	0.1092
We proposed	**0.1898**	**0.1465**	0.1322	0.1496	**0.1195**

(d) Average gradient comparison results of different algorithms.

Figure 11. Comparison of the results of each defogging algorithm in different environments.

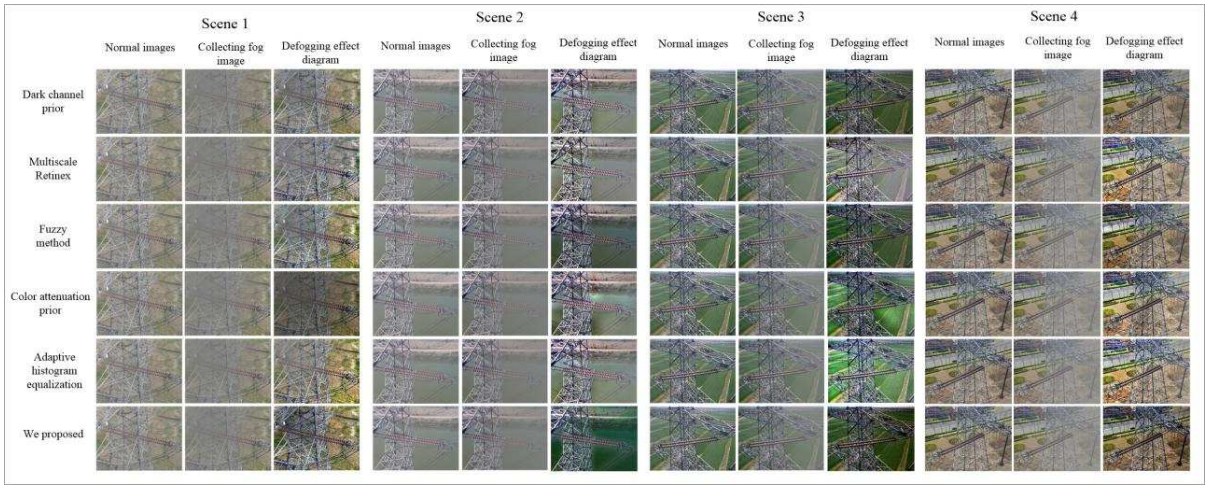


Figure 12. Visualization of execution effects of various algorithms in different environments.

From the analysis of the subjective visual effect of defogging restoration in Figure 12, the brightness, detail, color, and clarity of the restored image were greatly improved. However, this algorithm takes into account the difference of the depth of the target scene in the near and far scenes and can effectively overcome the problems of dim foreground and deepened shadow area. Among them, the haze removal effect by the color attenuation prior method in Scene 1 was poor; it made the foggy area form a high color saturation to block the target behind. In Scene 2, there is a certain degree of color distortion in the river part of the restored image, which is because the river part belongs to the high brightness area, because it does not meet the prior principle of dark channel. Scene 3 is a shot image with a grass background; the last three methods had a significant defogging effect, and the method proposed in this paper could restore a high contrast image. The Multiscale Retinex method had a general effect in such scenes. Scene 4 is based on the surrounding environment of an urban road; the proposed method performed well in the brightness, overall details, and contrast of the restored image. The overall analysis of visual effects shows that the improved dark channel prior defogging method can generate a restoration image with rich details and one close to the natural image, which can keep the overall structure of the image clear and highlight the edge details.

3.2. The Anchors Clustering Based on Improved K-Means and Genetic Algorithm Optimization

To get more suitable anchors for all kinds of targets in the experimental datasets, we used the $(1 - IOU)$ measurement method and the genetic algorithm to improve the original K-Means clustering algorithm. Nine groups of different anchors were sorted by size and replaced by the original ones. In the clustering process, the clusters were represented as nine color clusters. In the first stage of detection, three types of transmission line tower, insulator, and connected components with different sizes were clustered, and the visualization results are shown in Figure 13a. The results optimized by the $(1 - IOU)$ metric and the Genetic Algorithm are shown in Figure 13b. Six types of fittings, U-shackle, Triangular yoke plate, Adjustment plate, Tension clamp, Clamp bolt, and Insulator, with different sizes were clustered in the second-order detection datasets. The results are shown in Figure 13c, and the optimization results are shown in Figure 13d.

Figure 13 shows that each color cluster represents nine groups of anchors with different sizes (horizontal and vertical coordinates represent the width and height of anchors, respectively) in the clustering process. By iteratively updating the distance between each sample point and the cluster center, the distribution of each group of anchors was obtained as the reference value of the prediction box. After clustering optimization, the convergence speed of subsequent model training could be accelerated, and the average accuracy of the two optimized datasets was improved by 0.98%, which was better than the original

detection effect. The compared results between the proposed method and the original clustering identifies the best possible recall and fitness, as shown in Table 1.

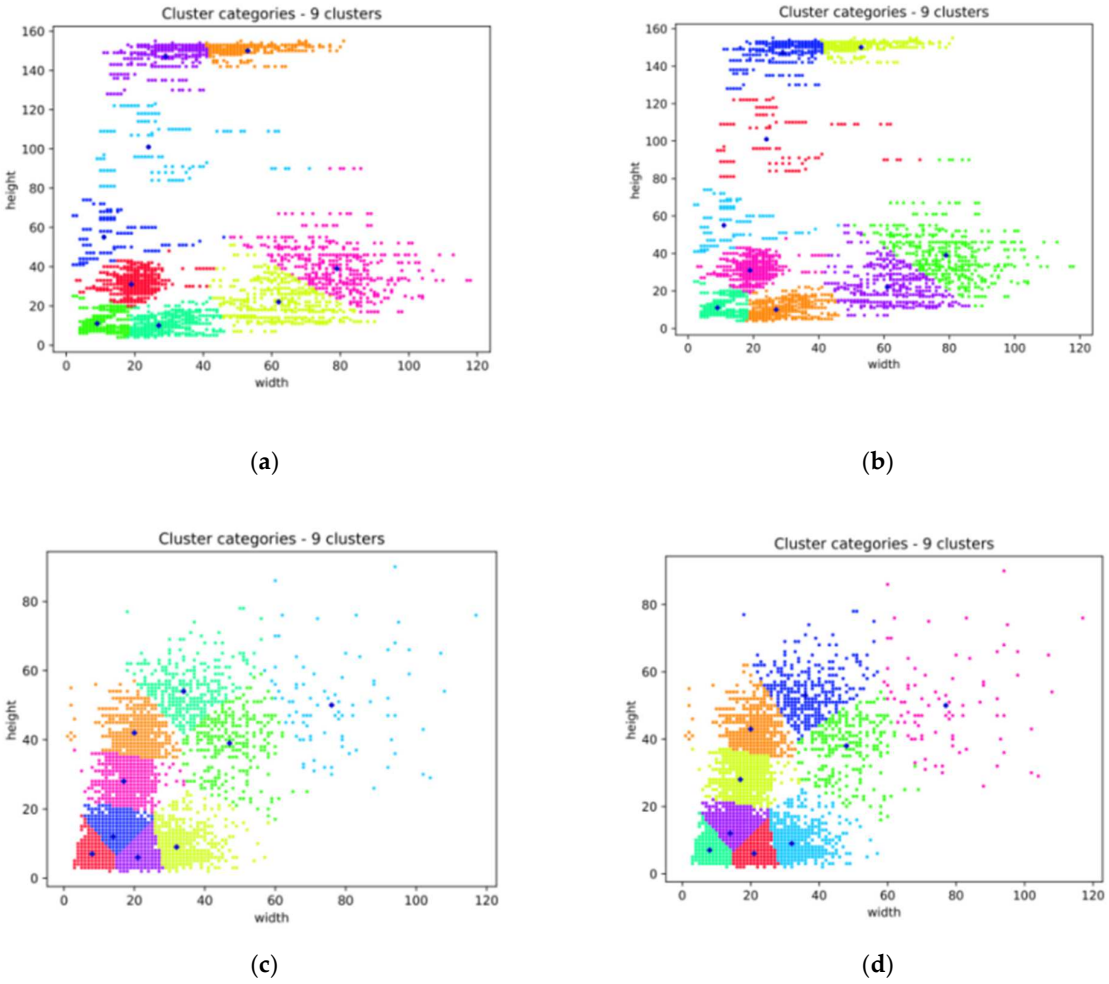


Figure 13. Visual comparison results of two-stage data clustering. (**a**) The results of one-stage anchors of K-means clustering; (**b**) The results of one-stage and genetic algorithm improvement; (**c**) The results of two-stage anchors of K-means clustering; (**d**) The results of two-stage anchors and genetic algorithm improvement.

Table 1. Comparison of clustering results of improved $(1 - IOU)$ and Genetic algorithm.

Stage	Methods	Best Possible Recall (%)	Fitness	Anchors
Stage one	Original K-means	0.99287	0.78492	[27,10] [11,55] [9,11] [62,22] [19,31] [79,39] [53,150] [29,147] [24,101]
	Our optimization	**1.00000 (+0.713)**	**0.78575 (+0.083)**	[25,11] [11,53] [10,11] [62,23] [19,30] [77,39] [51,151] [29,147] [22,100]
Stage two	Original K-means	0.99839	0.81036	[[8,7] [14,12] [21,6] [20,41] [33,9] [18,28] [76,50] [34,54] [45,38]
	Our optimization	**0.99951 (+0.112)**	**0.81751 (+0.715)**	[9,7] [15,13] [21,6] [21,41] [32,8] [18,29] [76,51] [34,55] [45,39]

Table 1 shows that the fitness and the best possible recall index of the two-stage datasets were optimized by the $(1 - IOU)$ metric method and Genetic Algorithm. The last column of the table shows the size of anchors, which adaptively matches various target sizes in the optimization process and makes correction and fine adjustment.

3.3. Model Performance Verification Experiment

3.3.1. Data Preprocessing

In this paper, a two-stage detection process was designed. Considering the dependence of Transformer on data sample size, to obtain the optimal experimental effect, we expanded the original datasets. The data enhancement methods included: flipping, rotation, clipping, scaling, color adjustment, erasing, MaxUp etc. The expanded one-stage detection datasets contained 6600 images, and the second-stage detection datasets contained 7700 images. In the first stage, the following three types of targets were detected by using the field collected datasets after defogging treatment: Transmission line tower, Insulator, and Connected components. The second stage intercepted the detection frame of connected components and detected the following six types of targets: U-shale (US), Triangular yoke plate (TYP), Adjustment plate (ADP), Tension clamp (TEC), Clamp bolt (CLB), and insulator (In). This experiment was carried out in the environment of the Ubuntu system, CUDA version is 10.1, GPU is NVIDIA GTX 1080TI, using Pytorch deep learning framework.

3.3.2. Two Stage Training and Testing Results

To verify the parameter performance of the Double-attention YOLO model, we conducted a series of comparative experiments under the same datasets and super parameters. We added different attention mechanism modules to the backbone part of the YOLOv5 model or replaced it with a series of lightweight pruning model networks. The comparison results of recall, precision, train loss, and val loss, in the first stage are shown in Figure 14.

In Figure 14a–d, the four attention network structures including SENet, Coordinate attention, Efficient channel attention and CBAM components were added to the backbone of YOLOv5, respectively. Figure 14a shows that recall performs best among all models with the attention mechanism, reaching 99.47%, and precision in Figure 14b was 99.89%, which was better than other models. The performance of the curve with the CBAM module improved slowly in the first 50 epochs, which was lower than other models, however, good accuracy was achieved in the later stage. The CBAM components were embedded in the backbone and neck part of YOLOv5 in our Double-attention YOLO model. The attention deployment on the whole network fully offsets the precision performance loss caused by the introduction of GhostNet model pruning and ViT module.

Figure 14c,d shows the training and validation loss function curves of several models. It shows that the overall decline process of loss function obtained by our method was smoother than other models, and the loss value reached the lowest at the 100th epoch, which were: train loss: 0.000621, Val loss: 0.000163. We also carried out comparative experiments on the network where the backbone of YOLOv5 was replaced by lightweight compression models; several types of lightweight models were used, EfficientNetLite, GhostNet, Mobilenet, PP-LCNet, and Sufflenetv2. Figure 14e,f shows that the recall and precision of the Double-attention YOLO model were 0.18% and 0.24% higher than those of the model whose backbone was only GhostNet, and 0.58% and 0.93% higher than other compression models in the best recall and precision performance. This shows that the global self-attention mechanism and channel space attention mechanism provided by Vision Transformer can make up for the precision loss caused by model compression when redundant parameters are eliminated. Figure 14g,h shows the compression model training and validation loss curve. It shows that the loss value of our method at the 100th epoch reached the global minimum, which was 0.000621 and 0.000163, respectively. Figure 14i–l shows the comparison results of recall and precision scores of various models in the second stage. It shows that the recall value of our method reached 0.9956, and the precision value

was 0.9983, which had the best effect among all the attention mechanisms and model compression backbone networks.

To further evaluate the detection accuracy of the Double-attention YOLO model and the performance of the network classifier, we drew a confusion matrix of the two-stage detection results (Figure 15). Figure 15a shows the classification accuracy of the three types of targets in the first stage and Figure 15b shows the classification accuracy of the six types of targets in the second stage. In the confusion matrix, the background region also participates in the performance evaluation as a category.

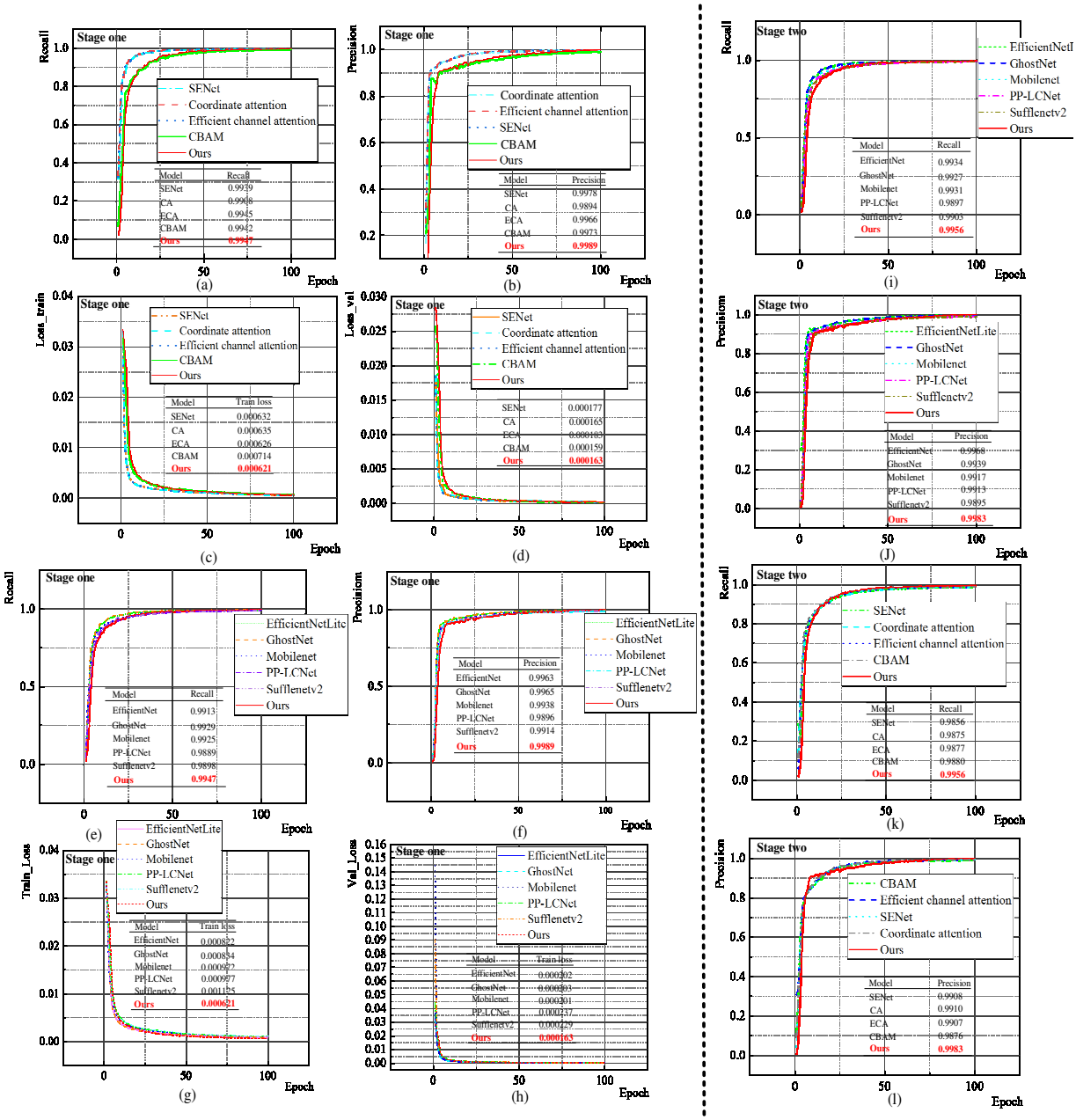
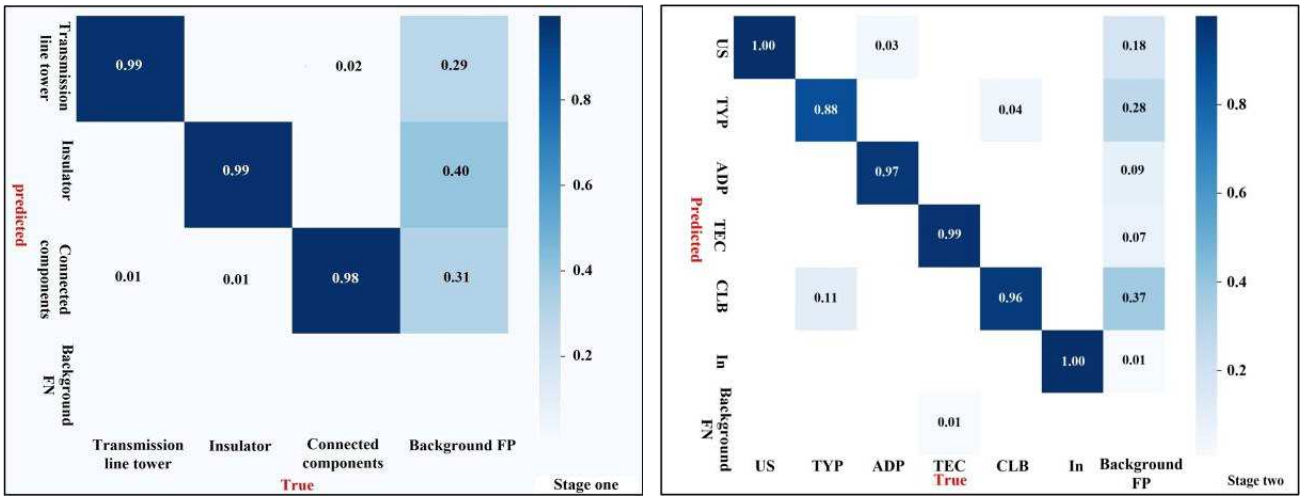


Figure 14. The first stage model detection performance comparison.



(**a**) (**b**)

Figure 15. Two-stage test sample confusion matrix. (**a**) The first stage test sample confusion matrix. (**b**) The second stage test sample confusion matrix.

According to the real value and prediction value of the confusion matrix, for normal and potential objects to be detected, the model classifier can generally reflect the high classification accuracy, but to a certain extent, there is a phenomenon of missing detection, which may be caused by serious target occlusion or difficult to capture small targets. In Figure 15a, the predicted values of the Transmission line tower and Insulator reach 99%, indicating that the network had good detection effect for such categories. Among them, the background misjudgment rates for Insulator and connected components were high, reaching 40% and 31%, respectively, indicating that the complex environmental background has interference factors for small targets. Inspection of Figure 15b shows that the predictive value for the Triangular yoke plate (TYP) was relatively low and there was a probability that it would be misjudged as the Clamp bolt (CLB), because many Clamp bolts were connected to the Triangular yoke plate. However, the initial annotation box of the two categories may overlap, resulting in a small number of errors in the generation of anchors. The prediction accuracy of other categories such as U-shackle (US), Insulator (In), Tension clamp (TEC), and so on, were all above 99%. The detection accuracy and classification effect in the second stage were excellent.

To verify the dual roles of the global multi-head self-attention networks proposed in the Double-attention YOLO model in enhancing global receptive field and channel spatial attention mechanism in capturing local information, we visualized the attention degree of the model by thermal map. The original image is shown in Figure 16a, and the imaging effect of the baseline model YOLOv5 is shown in Figure 16b, the effect picture of our method is shown in Figure 16c.

Figure 16 shows that the red and some color regions in the graph are the parts of the network model that are of particular concern. The diagram contains six categories: US, TYP, ADP, TEC, CLB and metal rust areas. The darker the color, the higher the visual saliency. Compared with the original baseline model, the global multi-head self-attention network in Vision Transformer can provide more feature information for the network input, and the global receptive field of the model is enhanced by global search of feature map, while the saliency of the target to be detected in a complex background environment can be enhanced by CBAM module.

Figure 17 shows that after extracting the shallow feature information, the model can effectively segment the foreground region and background region where the defect is located. After pruning using the GhostNet module linear transformation, the down sampling effect of feature map is obviously enhanced. Nonetheless, the ViT module provides more

abstract global spatial information for the feature graph, and the semantic information of the high-level of the feature graph is fuzzy. Finally, through the CBAM module, the feature extraction network can accurately capture the local fine-grained information. From the visualization results of feature map, we can see that the Double-attention YOLO model can fully extract the texture, shape, and edge information in the datasets image of transmission line connection fittings.

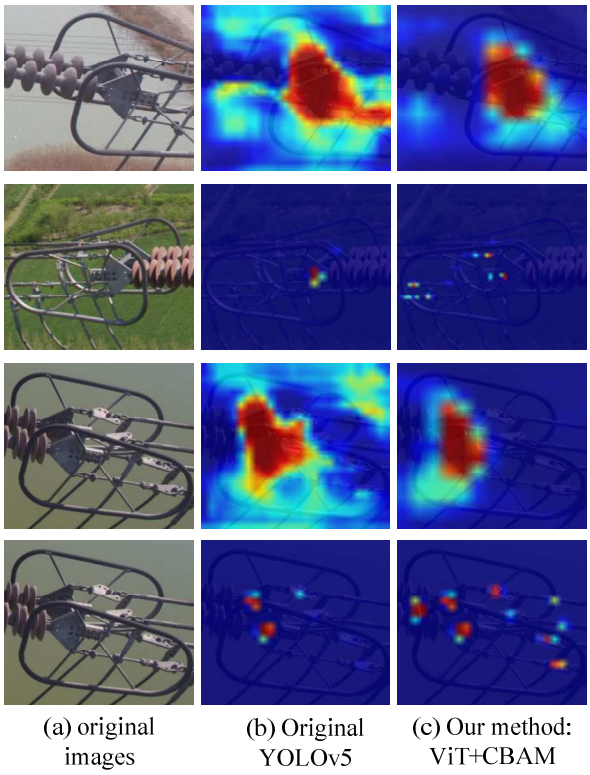


Figure 16. Visualization results of thermal diagram of model attention.

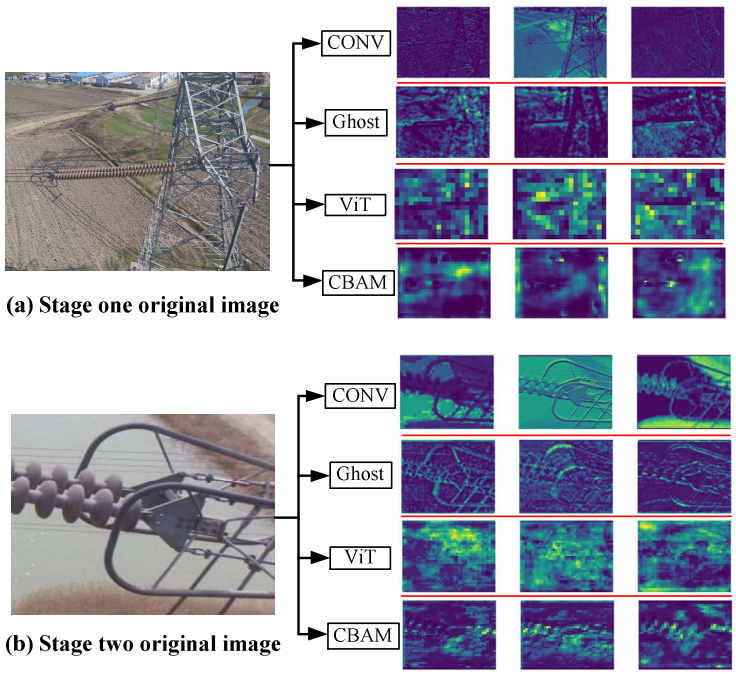


Figure 17. Double-attention YOLO network main module feature extraction visualization results.

To further evaluate the classification performance of the model, the classification accuracy rate P, R, and macro-F1 were counted, respectively. Macro-F1 combines the harmonic average value of P and R to improve the precision and recall and reduce the difference between them as far as possible. Macro-$F_1 \in [0, 1]$, and the formulae for P, R, and macro-F1 are shown in Formulas (18) and (19):

$$P_i = \frac{TP_i}{TP_i + FP_i}, R_i = \frac{TP_i}{TP_i + FN_i}, P_{macro} = \frac{\sum_{i=1}^{n} P_i}{n}, R_{macro} = \frac{\sum_{i=1}^{n} R_i}{n} \tag{18}$$

$$F1_{macro} = \frac{2 \times P_{macro} \times R_{macro}}{P_{macro} + R_{macro}} \tag{19}$$

Among them, TP_i to predict the correct number of positive samples; FP_i is the number of positive samples with prediction errors; TN_i is the number of positive samples was negative; FN_i to predict the number of negative samples with errors. The classification accuracy, recall rate, and $F1_{macro}$ values of each category are calculated by Formula (19). The statistical results of $F1_{macro}$ and the classification evaluation index are shown in Figure 18, in which Figure 18a shows the test results of the first stage, and Figure 18b shows the test results of the second stage.

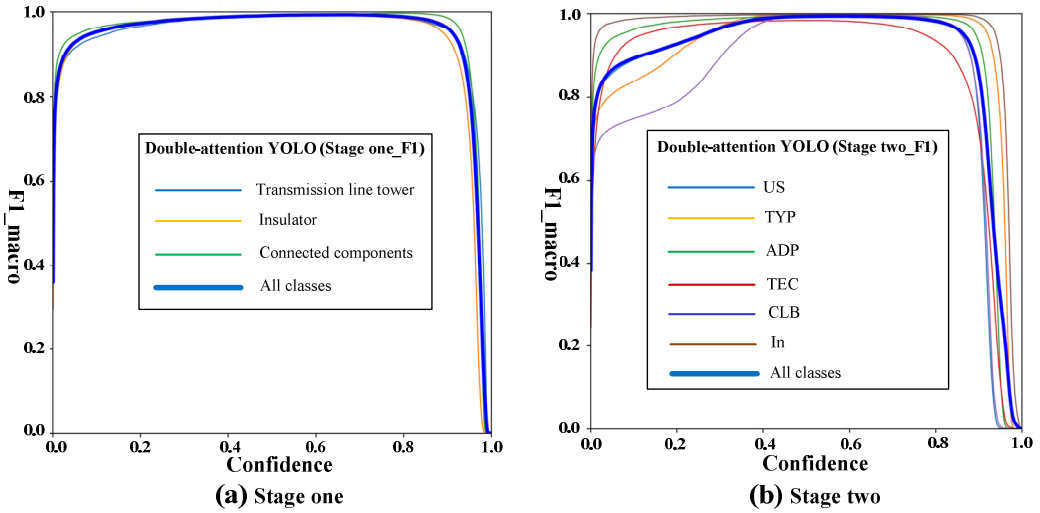


Figure 18. The results of macro-F1.

The abscissa in Figure 18 is the confidence level of each category, and the dark bold curve is the weighted statistical result for all categories. The score curve shows that $F1_{macro}$ finds the optimal balance between Precision and Recall. In the first stage, the optimal balance threshold for precision and recall is 0.661 for all categories in Figure 18a, and 0.568 for all categories in Figure 18b, and the Precision and Recall scores of the model are optimal.

To comprehensively evaluate the effect of the model, we used several different attention mechanism modules and the model compression network as the replacement components of the model backbone of YOLOv5 and verified the results obtained. As shown in Tables 2 and 3, we compared parameters, FLOPs, AR, mAP@0.5:0.95, FPS, and the best results are displayed in bold font. To ensure the objective fairness of the results, all the comparative experiments were set with the same super parameters, and the training and testing were carried out under a unified framework.

Table 2. Verification results under different backbone components (Stage one).

Model	Backbone	Params	FLOPs	AR	mAP@0.5:0.95	FPS
YOLOv5s (Baseline)	CSP-Darknet53	7.0 M	15.8 G	0.991	0.9023	93
YOLOv5s-SENet	CSPv5+SE block	7.2 M	16.6 G	0.994	0.9406	96
YOLOv5s-CA	CSPv5+Coordinate attention block	7.13 M	16.3 G	0.992	0.9418	91
YOLOv5s-ECA	CSPv5+Efficient channel attention	7.08 M	16.5 G	0.990	0.9113	93
YOLOv5s-CBAM_backbone	CSPv5+CBAM block	7.1 M	16.8 G	0.995	0.9425	90
YOLOv5m-Sufflenetv2_backbone	ShuffleNetv2	21.2 M	40.4 G	0.988	0.8812	61
YOLOv5m-PP-LCNet_backbone	PP-LCNet	21.6 M	41.5 G	0.988	0.8796	79
YOLOv5m-Mobilenetv3Small_backbone	MobileNet_v3Small	20.3 M	38.2 G	0.992	0.8909	67
YOLOv5m-EfficientNetLite_backbone	EfficientNet_Lite	22.9 M	43.8 G	0.990	0.9144	65
YOLOv5m-GhostNet_backbone	GhostNet	24.3 M	42.3 G	0.994	0.9187	63
YOLOv5m-Swin_Transformer _backbone	CSPv5+Swin Transformer block	102.3 M	225.9 G	0.989	0.9464	19
Ours	CSPv5+ViT+CBAM block	112.9 M	270.1 G	**0.996 (0.3% ↑)**	**0.9480 (4.6% ↑)**	20

Table 3. Verification results under different backbone components (Stage two).

Model	Backbone	Params	FLOPs	AR	mAP@0.5:0.95	FPS
YOLOv5m (Baseline)	CSP-Darknet53	21.2 M	49.2 G	0.986	0.8241	79
YOLOv5m-SENet	CSPv5+SE block	21.3 M	48.7 G	0.986	0.8241	84
YOLOv5m-CA	CSPv5+Coordinate attention block	21.3 M	48.5 G	0.987	0.8246	83
YOLOv5m-ECA	CSPv5+Efficient channel attention	21.3 M	48.6 G	0.987	0.8239	84
YOLOv5m-CBAM_backbone	CSPv5+CBAM block	21.3 M	48.7 G	0.989	0.8264	79
YOLOv5m-Sufflenetv2_backbone	ShuffleNetv2	21.2 M	40.4 G	0.987	0.8432	62
YOLOv5m-PP-LCNet_backbone	PP-LCNet	21.6 M	41.5 G	0.988	0.8346	80
YOLOv5m-Mobilenetv3Small_backbone	MobileNet_v3Small	20.3 M	38.3 G	0.993	0.8324	65
YOLOv5m-EfficientNetLite_backbone	EfficientNet_Lite	22.8 M	43.8 G	0.993	0.8432	58
YOLOv5m-GhostNet_backbone	GhostNet	24.2 M	42.3 G	0.992	0.8474	53
YOLOv5m-Swin_Transformer _backbone	CSPv5+Swin Transformer block	102.3 M	225.2 G	0.982	0.8337	15
Ours	CSPv5+ViT+CBAM block	112.9 M	270.3 G	**0.994 (0.8% ↑)**	**0.8674 (4.3% ↑)**	16

In Tables 2 and 3, the first two columns list the names of the various models and their corresponding backbone main structures, and the third to seventh columns are the performance evaluation indexes of the various models. The two-stage independent detection results are presented in Tables 2 and 3.

Tables 2 and 3 show that the CBAM achieves the best effect of 0.995 in the replacement results of the attention mechanism module, and GhostNet achieves the best result of 0.994 in the model compression replacement results, while our method is the best among all the replacement results of backbones, reaching 0.996, which is 0.3 percentage points ahead of the baseline model.

In the second stage, the AR score of CBAM and EfficientNet in the replacement component results achieved the best effect in their respective groups, 0.989 and 0.993, respectively. Our method was 0.994, which was ahead of the baseline model by 0.8 percentage points. We adopted a more rigorous approach, mAP@0.5:0.95, as the average accuracy index, the CBAM and GhostNet replacement components obtained the best scores of 0.9425 and 0.9187, respectively, in their groups, and our method reached 0.9480, which was 4.6 percentage points ahead of the baseline model, which was the highest overall level. In the score of the

second stage test, the Coordinated attention block and GhostNet achieved the best results of 0.8246 and 0.8624, respectively, in their groups, while our method still obtained the highest overall score of 0.8674, 4.3 percentage points ahead of the baseline model. Replacement of the GhostNet component in the mAP@0.5:0.95 had more advantages in the comprehensive effect, and the floating-point computation per second was 42.3 G, which was better than the baseline model and other lightweight models. Our model had the same parameters and FLOPs as the YOLO framework with backbone as the Swin Transformer, reaching 112.9 M and 270 G. The model reasoning speed was also above that for the YOLOv5m-Swin_Transformer_backbone, and the FPS was 20 and 16 in the two phases. Although the introduction of the Transformer block results in too much calculation and too many model parameters, the method improved the performance of various precision parameters, and the potential application value of Vision Transformer in visual tasks such as target detection was verified, which may be better than the traditional mainstream CNN model. This fully demonstrated that the Double-attention YOLO network can not only enhance the global receptive field of the model, but also capture the local salient features. It can fully integrate the advantages of the dual attention mechanism, to solve the problem of detecting small targets with large scene depth and overlapping targets in a complex environment.

To evaluate the ability of our proposed method to deal with a complex detection environment more objectively, six kinds of current advanced target detection networks were selected for further comparison. The selected detection networks were ATSS, Faster RCNN+FPN, FCOS, SSD, RetinaNet, and Deformable DETR. The Deformable DETR processes the feature map through multi-scale attention module, for the current more advanced target detection Transformer framework. We also added the three baseline improvement models of YOLOv5m-CBAM_backbone, YOLOv5m-GhostNet_ backbone, and YOLOv5m-Swin Transformer _ backbone to the comparative experiment. In particular, we applied the current advanced Swin Transformer module based on the sliding window detection to the backbone of the YOLOv5 baseline model, and further explored the potential advantages of Transformer in visual tasks. To ensure objective fairness, we conducted unified training on 7700 second stage target detection datasets, and used the same index to verify the results mAP@0.5, mAP@0.75, mAP@0.5:0.95 (Table 4).

Table 4. Comparison of test performance of various algorithms.

Framework	Backbone	mAP@0.5	mAP@0.75	mAP@0.5:0.95
ATSS	ResNet50	0.988	0.945	0.846
Faster RCNN+FPN	ResNet50	0.949	0.767	0.672
FCOS	ResNet50	0.970	0.795	0.685
SSD	VGG16	0.969	0.826	0.699
RetinaNet	ResNet50	0.885	0.566	0.531
Deformable DETR	ResNet50	0.976	0.767	0.661
YOLOv5m-CBAM_backbone	CSPv5+CBAM block	0.9941	0.904	0.8264
YOLOv5m-GhostNet_backbone	GhostNet	0.9937	0.9108	0.8374
YOLOv5m-Swin Transformer _backbone	CSPv5+Swin Transformer block	0.9940	0.9177	0.8337
Ours	CSPv5+ViT +CBAM block	**0.9948**	0.9302	**0.8674**

Table 4 shows that our Double-attention YOLO model achieved high scores in three performance metrics, among them, the average precision of mAP@0.5:0.95 was 3.37 percentage points higher than the advanced YOLOv5m-Swin Transformer _ backbone, reaching 86.74%, while the ATSS achieved an average accuracy of 94.5% on mAP@0.75 through adaptive sample training, leading our approach. However, the score of the Double-attention YOLO on mAP@0.5 was also ahead of all algorithms, reaching 99.48%. This also showed that our method effectively combined the global self-attention function of Vision Transformer based on GhostNet model pruning and the local feature saliency capture function of CBAM and has great advantages in three kinds of average accuracy evaluation indexes.

To further verify the effect of our proposed Double-attention YOLO model in dealing with large scene depth, small target detection, and target occlusion, three kinds of

new images without network training were selected to verify the model detection effect. The selected test data had the following characteristics: the background conditions were complex and diverse, there were mixed multi-scale targets to be detected, as shown in Figure 19. Figure 19a–e shows the detection result of the first stage, Figure 19f–k shows the detection result of the second stage, and Figure 19k shows the rust fault detection results of metal parts in the second stage. In this test, CBAM and GhostNet were used to replace the backbone of the YOLOv5 baseline model, and the detection results of the Double-attention YOLO model were compared. Therefore, it reflects the detection advantages of this paper after integrating the above model structures. The following three types of models were called model 1, model 2, and model 3, where the first line of each phase enters the original image of the test, the second line manually marks the ground truth, followed by the test result.

(a) Original images (Stage one).

(b) Ground truth (Stage one).

(c) The detection results of YOLOv5m-CBAM_backbone (Stage one).

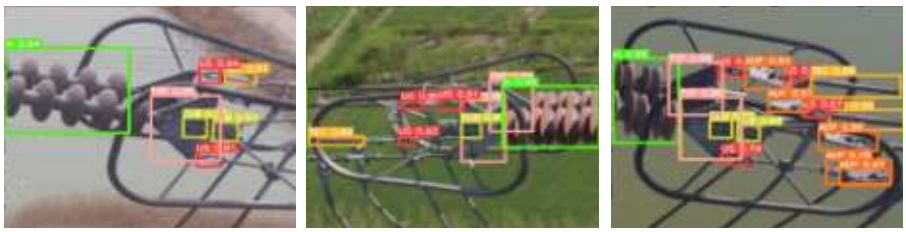
(d) The detection results of YOLOv5m-GhostNet_backbone (Stage one).

(e) The detection results of us (Stage one).

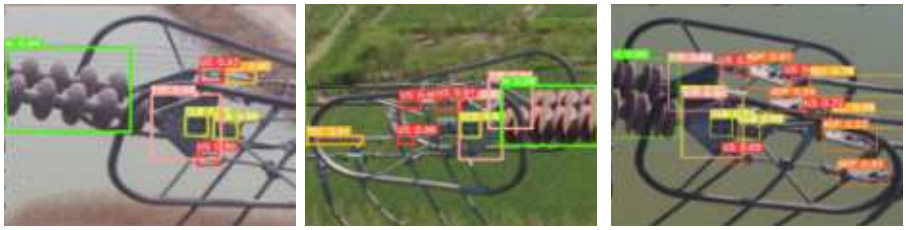
Figure 19. *Cont.*

(f) Original images (Stage two).

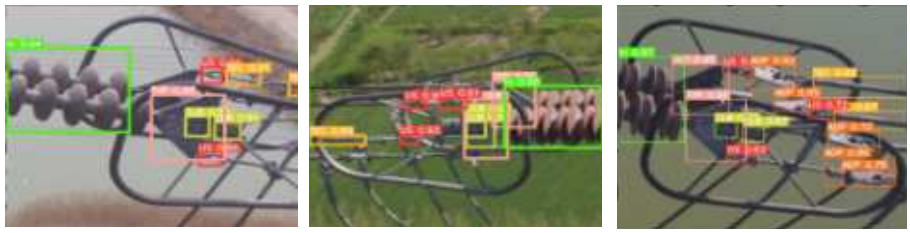
(g) Ground truth (Stage two).



(h) The detection results of YOLOv5m-CBAM_backbone (Stage two).



(i) The detection results of YOLOv5m-GhostNet_backbone (Stage two).



(j) The detection results of us (Stage two).

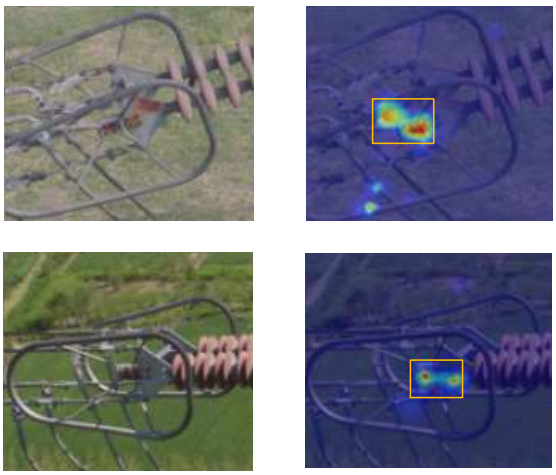
(k) The detection results of rust components (Stage two).

Figure 19. Model detection results. (**a**–**e**) stage one. (**f**–**k**) stage two.

The results show that the first stage is the detection of complex environment targets in a large scene depth. There are three types of detection objects. Because the target to be tested was confused with the background and the color was similar, model 1 (Figure 19c)

missed an insulator string under a paddy field and land background, while the recognition rate of model 2 (Figure 19d) was only 0.33. The proposed method (Figure 19e) had a recognition rate of 0.73, and in the recognition accuracy of the Transmission line tower and Connected components, our method (Figure 19e) had higher recognition accuracy, which shows that the neural network can capture more useful features by global self-attention enhancement. The second stage was the detection of small targets and dense targets and there was a certain degree of object occlusion, as shown in Figure 19f–k. The prediction box and the ground truth of the output for the detection results in the three models were very high and there were only individual missed cases, indicating that our method can capture the key information to be detected in a complex scene and has a strong feature extraction capability adapted to this dataset. Model 1 (Figure 19h) and model 2 (Figure 19i) missed detection of the Tension clamp (TEC), Clamp bolt (CLB), and Adjustment plate (ADP), respectively. These missed targets had the characteristics of small size and occlusion interference. At the same time, our method (Figure 19j), had more advantages in the recognition accuracy of other types of connecting fittings. This fully shows that the combination of the ViT module and the CBAM module can consider the global and local saliency of the image, so that the detection network can better cope with various complex environmental conditions. This shows that the recognition accuracy of the corrosion part is about 90%. In addition, in Figure 19k, we performed routine rust detection on the connecting fittings with different degrees of rust faults. The analysis of the detection results shows that the proposed algorithm can achieve good results in the case of identifying some rust faults. The average detection accuracy in the example was maintained at about 80% and the highest score was 95%.

We introduced the CBAM attention mechanism into the network structure of the Double-attention YOLO to enhance the saliency of rusty regions in complex backgrounds, which was used to improve the recognition accuracy of the model for variable rusty regions. Figure 20a,b shows two images with rust defects that were randomly selected and the local attention heat map after introducing the attention mechanism was visualized. Figure 20a is the original image and Figure 20b is the thermal region map that the attention mechanism focused on.



(a) Original rust defect image

(b) Attention figure of rusting part

Figure 20. Visualization results of local attention map in rust area.

Figure 20 shows that the darker the heat map area in the attention map, the greater is the probability of rust defects. The CBAM attention mechanism can effectively locate the rust area to be detected, further reducing the focus of model feature extraction. The part, and the CBAM module can focus on the rust area with high uncertainty of area, shape, and

color in a complex environment, thereby helping the model to improve the efficiency of feature extraction.

3.3.3. Design of Condition Monitoring System for Transmission Line Connection Fittings

In the process of the algorithm design, we also deployed the hardware equipment and the detection algorithm and developed the condition monitoring system of transmission line connecting fittings based on mobile terminal application. The software interface for the application system is shown in Figure 21. Its functions include calling the camera for real-time state monitoring, outputting the algorithm identification results of each stage, judging the fault type and warning. Users can adjust the state parameters of the UHV transmission line according to the actual situation to adapt to different application scenarios. The data acquisition sources of this software interface were the tower end HD camera and the line inspection robot, as shown in Figure 2a,b. Therefore, the installation location of the equipment, the acquisition angle of the image, the construction of the data analysis platform, and the connection of the LAN need to be properly adjusted according to the field environment.

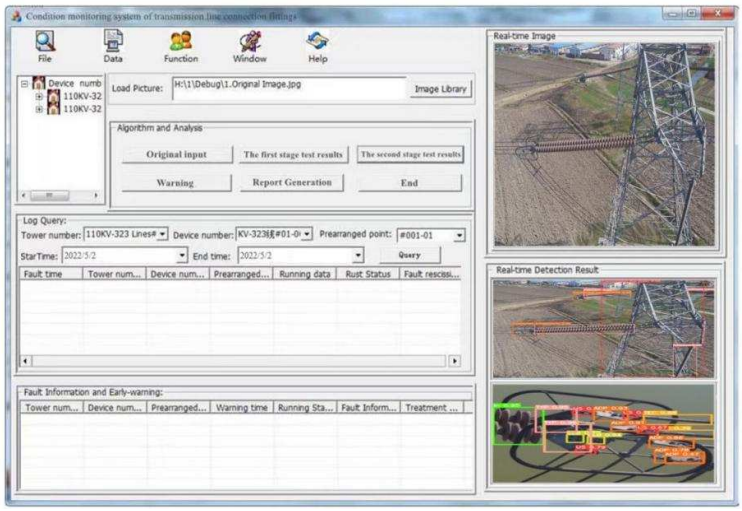


Figure 21. Condition monitoring system of transmission line connection fittings.

4. Conclusions

In this paper, an improved dark channel prior defogging algorithm is proposed to preprocess the images of transmission line fittings taken in complex environments to obtain high quality available data. Based on the YOLOv5 model, through the integration of Vision Transformer, a channel and spatial attention mechanism, and the GhostNet model compression module, we built a new model with global image perception and local saliency capture ability: the Double-attention YOLO. The model detects small targets in large scene depth through a hybrid mechanism of dual attention and can deal with dense prediction tasks with occlusion interference. The experimental results showed that our proposed method is superior to all the advanced baseline improved YOLOv5 models, which are replaced by attention mechanisms and model compression units and is superior to the current advanced one-stage, two-stage and the Vision Transformer target detection methods in terms of map index determined by IOU threshold in three ranges. The main contributions of this paper are as follows:

1. An improved dark channel prior defogging algorithm is proposed to solve the preprocessing problem of transmission line fittings in complex environments.
2. The potential advantages and application value of the multi-head self-attention mechanism in Vision Transformer for dense target prediction tasks are verified.
3. In the data preprocessing stage, the advanced MixUp data enhancement strategy is adopted, and the feature fusion of multi-scale targets is realized through the BiFPN

structure. The optimal anchors are generated by improved the K-Means clustering algorithm and the genetic algorithm to match the diversity of target information in the datasets.

4. In YOLOv5, Vision Transformer, channel and spatial attention mechanism, and the GhostNet model compression unit are integrated. Compared with the original baseline model and the improved baseline model, the performance of YOLOv5 is greatly improved, and it is better than several current advanced target detection algorithms.

Author Contributions: All authors contributed to the study conception and design; preliminary experiment, methodology, writing–original draft, Z.S.; verification, writing–review and editing, supervision, X.H. and C.J.; project management and collecting documents, modifying formats, reference materials, Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Natural Science Basis Research Plan in Shaanxi Province of China (Program No. 2022JQ-568); the Scientific Research Program Funded by Shaanxi Provincial Education Department (Program No. 21JK0661); the Key Research and Development Projects in Shaanxi Province (2021GY-306); and the Key R&D plan of Shannxi (2021GY-320, 2020ZDLGY09-10).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors have no financial or proprietary interests in any material discussed in this article.

References

1. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 23–28 June 2014; pp. 580–587.
2. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, USA, 7–12 December 2015; Volume 28.
3. Wu, W.; Yin, Y.; Wang, X.; Xu, D. Face Detection with Different Scales Based on Faster R-CNN. *IEEE Trans. Cybern.* **2015**, *49*, 4017–4028. [CrossRef] [PubMed]
4. Mai, X.; Zhang, H.; Jia, X.; Meng, M.Q. Faster R-CNN With Classifier Fusion for Automatic Detection of Small Fruits. *IEEE Trans. Autom. Sci. Eng.* **2021**, *17*, 1555–1569. [CrossRef]
5. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland, 2016; pp. 21–37.
6. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
7. Bochkovskiy, A.; Wang, C.Y.; Liao HY, M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
8. Lu, X.; Ji, J.; Xing, Z.; Miao, Q. Attention and Feature Fusion SSD for Remote Sensing Object Detection. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 5501309. [CrossRef]
9. Ge, H.; Dai, Y.; Zhu, Z.; Liu, R. A Deep Learning Model Applied to Optical Image Target Detection and Recognition for the Identification of Underwater Biostructures. *Machines* **2022**, *10*, 809. [CrossRef]
10. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
11. Zaremba, W.; Sutskever, I.; Vinyals, O. Recurrent neural network regularization. *arXiv* **2014**, arXiv:1409.2329.
12. Graves, A. Long short-term memory. In *Supervised Sequence Labelling with Recurrent Neural Networks*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 37–45.
13. He, J.; Zhao, L.; Yang, H.; Zhang, M.; Li, W. HSI-BERT: Hyperspectral image classification using the bidirectional encoder representation from transformers. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 165–178. [CrossRef]
14. Sung, C.; Dhamecha, T.I.; Mukhi, N. Improving short answer grading using transformer-based pre-training. In Proceedings of the International Conference on Artificial Intelligence in Education, Chicago, IL, USA, 25–29 June 2019; Springer: Cham, Switzerland, 2019; pp. 469–481.
15. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving language understanding by generative pre-training. *J. Softw. Eng. Appl.* **2018**, *12*, 11.
16. Chen, Y.; Zhang, X.; Chen, W.; Li, Y.; Wang, J. Research on Recognition of Fly Species Based on Improved RetinaNet and CBAM. *IEEE Access* **2020**, *8*, 102907–102919. [CrossRef]

17. Wei, S.; Qu, Q.; Su, H.; Shi, J.; Zeng, X.; Hao, X. Intra-pulse modulation radar signal recognition based on Squeeze-and-Excitation networks. *Signal Image Video Process.* **2020**, *14*, 1133–1141. [CrossRef]
18. Jiang, G.; Jiang, X.; Fang, Z.; Chen, S. An efficient attention module for 3d convolutional neural networks in action recognition. *Appl. Intell.* **2021**, *51*, 7043–7057. [CrossRef]
19. Xie, J.; Miao, Q.; Liu, R.; Xin, W.; Tang, L.; Zhong, S.; Gao, X. Attention adjacency matrix based graph convolutional networks for skeleton-based action recognition. *Neurocomputing* **2021**, *440*, 230–239. [CrossRef]
20. Paoletti, M.E.; Haut, J.M.; Pereira, N.S.; Plaza, J.; Plaza, A. Ghostnet for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 10378–10393. [CrossRef]
21. Yue, X.; Li, H.; Shimizu, M.; Kawamura, S.; Meng, L. YOLO-GD: A Deep Learning-Based Object Detection Algorithm for Empty-Dish Recycling Robots. *Machines* **2022**, *10*, 294. [CrossRef]
22. Du, F.; Jiao, S.; Chu, K. Research on Safety Detection of Transmission Line Disaster Prevention Based on Improved Lightweight Convolutional Neural Network. *Machines* **2022**, *10*, 588. [CrossRef]
23. Yan, S.; Chen, P.; Liang, S.; Zhang, L.; Li, X. Target Detection in Infrared Image of Transmission Line Based on Faster-RCNN. In Proceedings of the International Conference on Advanced Data Mining and Applications, Sydney, NSW, Australia, 2–4 February 2022; Springer: Cham, Switzerland, 2022; pp. 276–287.
24. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. Ghostnet: More features from cheap operations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 1580–1589.
25. Wang, B.; Chen, S.; Wang, J.; Hu, X. Residual feature pyramid networks for salient object detection. *Vis. Comput.* **2020**, *36*, 1897–1908. [CrossRef]
26. Yang, J.; Fu, X.; Hu, Y.; Huang, Y.; Ding, X.; Paisley, J. PanNet: A deep network architecture for pan-sharpening. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5449–5457.
27. Chen, J.; Mai, H.; Luo, L.; Chen, X.; Wu, K. Effective feature fusion network in BIFPN for small object detection. In Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 19–22 September 2021; pp. 699–703.
28. Zhu, Q.; Mai, J.; Shao, L. A Fast Single Image Haze Removal Algorithm Using Color Attenuation Prior. *IEEE Trans. Image Process.* **2015**, *24*, 3522–3533. [CrossRef]
29. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv* **2017**, arXiv:1710.09412.
30. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
31. Wang, C.Y.; Liao HY, M.; Wu, Y.H.; Chen, P.Y.; Hsieh, J.W.; Yeh, I.H. CSPNet: A new backbone that can enhance learning capability of CNN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 390–391.
32. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
33. Ottakath, N.; Al-Ali, A.; Al Maadeed, S. Vehicle identification using optimised ALPR. In Proceedings of the Qatar University Annual Research Forum and Exhibition (QUARFE 2021), Doha, Qatar, 20 October 2021.
34. Ding, L.; Goshtasby, A. On the Canny edge detector. *Pattern Recognit.* **2001**, *34*, 721–725. [CrossRef]
35. Al-Rahlawee AT, H.; Rahebi, J. Multilevel thresholding of images with improved Otsu thresholding by black widow optimization algorithm. *Multimed. Tools Appl.* **2021**, *80*, 28217–28243. [CrossRef]
36. He, K.; Sun, J.; Tang, X. Single image haze removal using dark channel prior. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *33*, 2341–2353. [PubMed]
37. Tufail, Z.; Khurshid, K.; Salman, A.; Nizami, I.F.; Khurshid, K.; Jeon, B. Improved Dark Channel Prior for Image Defogging Using RGB and YCbCr Color Space. *IEEE Access* **2018**, *6*, 32576–32587. [CrossRef]
38. Fan, T.; Li, C.; Ma, X.; Chen, Z.; Zhang, X.; Chen, L. An improved single image defogging method based on Retinex. In Proceedings of the International Conference on Image, Vision and Computing (ICIVC), Chengdu, China, 2–4 June 2017; pp. 410–413. [CrossRef]
39. Koley, S.; Sadhu, A.; Roy, H.; Dhar, S. Single Image Visibility Restoration Using Dark Channel Prior and Fuzzy Logic. In Proceedings of the International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech), Kolkata, India, 4–5 May 2018; pp. 1–7.
40. Kapoor, R.; Gupta, R.; Son, L.H.; Kumar, R.; Jha, S. Fog removal in images using improved dark channel prior and contrast limited adaptive histogram equalization. *Multimed. Tools Appl.* **2019**, *78*, 23281–23307. [CrossRef]