applied sciences

MDPI

Article

An Improved YOLOv7 Model Based on Visual Attention Fusion: Application to the Recognition of Bouncing Locks in Substation Power Cabinets

Yang Wang [1], Xiaofeng Zhang [1], Longmei Li [1,*], Liming Wang [1], Ziyang Zhou [1] and Peng Zhang [2]

[1] School of Electrical Engineering, Naval University of Engineering, Wuhan 430030, China; 18726996098@163.com (Y.W.); zhouziyang0969@163.com (X.Z.); licesoar@163.com (L.W.); m22385403@nue.edu.cn (Z.Z.)
[2] The 92808th Unit of the People's Liberation Army, Sanya 572000, China; 15926325989@163.com
* Correspondence: vivianlee527@163.com

Featured Application: This work can be applied to the task of environmental perception of a mobile manipulator based on visual guidance, such as opening a spring lock to open a door in the dangerous environment of a substation.

Abstract: With the continuous progress of intelligent power system technology, in order to meet the needs of substation operation and maintenance, a target detection algorithm is applied to identify the status of equipment switches. YOLOv7, as the latest achievement of YOLO (You Only Look Once) series algorithms, has good speed and accuracy in target detection tasks. However, when the generalized network is applied in a specific scenario, its advantages are not obvious due to its high weight and poor portability. In this paper, an improved GF-YOLOv7 network model is proposed to apply in the recognition of the status of bounce locks in a substation. The MobileViT module is used to improve the feature extraction ability of the backbone network. Referring to the CBAM feature attention mechanism, the channel attention module and the spatial attention module are used to design a more lightweight feature fusion network. The experimental results in the test set show that the proposed network can significantly reduce the network weight and improve the detection accuracy on the basis of a small reduction in the detection speed, and the accuracy reaches 97.8%, which can meet the needs of the detection task of substation bounce locks.

Keywords: substation bounce lock; YOLOv7; attention mechanism; target detection

Citation: Wang, Y.; Zhang, X.; Li, L.; Wang, L.; Zhou, Z.; Zhang, P. An Improved YOLOv7 Model Based on Visual Attention Fusion: Application to the Recognition of Bouncing Locks in Substation Power Cabinets. *Appl. Sci.* **2023**, *13*, 6817. https://doi.org/10.3390/app13116817

Academic Editor: Stéfano Frizzo Stefenon

Received: 26 April 2023
Revised: 23 May 2023
Accepted: 29 May 2023
Published: 4 June 2023

1. Introduction

In substations, condition testing of electrical equipment locks is an important part of maintaining the safety of equipment and personnel. The failure of electrical equipment locks can lead to threats to personnel safety and equipment [1,2]. Therefore, the status of electrical equipment locks needs to be tested regularly to detect problems and take corresponding measures in time to ensure the normal operation of equipment and the safety of personnel in substations. In order to accurately detect the lock status of electrical equipment, researchers use image detection [3], thermal imaging [4], and infrared temperature measurement [5,6] to quickly and accurately detect lock faults. These methods can detect the lock status, which is difficult to distinguish with the naked eye, and prevent the occurrence of electrical equipment malfunction. In several studies, experts have confirmed the importance of substation electrical equipment lock condition detection. For example, Zheng et al. [7] proposed an infrared insulator image detection model based on improved feature fusion for single-shot multi-box detectors to evaluate substation electrical equipment. Ciric et al. [8] highlighted the advantages of thermography in monitoring the status of substation electrical cabinet door locks. Wang et al. [9] proposed a deep learning-based

substation switchgear status recognition method based on the YOLOv3 network. Therefore, the detection of electrical equipment lock status is of extreme importance to ensure the safety and reliability of substation equipment through preventive maintenance.

The target detection algorithm is an important part of computer vision. It can obtain image features by calculating the weight parameters to discriminate the location and type of the target to be detected in the image. Common target detection models are region-based CNN (Convolutional Neural Network) [10,11], SSD (Single Shot Detector) series [12], YOLO (You Only Look Once) series [13,14], DETR (Detection Transformer) series [15], and their derived networks [16,17]. The two-stage algorithm represented by R-CNN and its derived network generally performs the target detection process as follows: (1) extract possible areas with targets by using a filtering algorithm; (2) obtain features by the CNN operation to determine the presence and category of targets in the candidate area; (3) calculate the coordinates of the position frame of the returned object by convolution to improve the positioning accuracy. Fast R-CNN changes the way R-CNN extracts the related features of each candidate box from one operation; it then scales each area of the feature map using the pooled features of the area of interest, and finally classifies and regresses the features and locations of the target [18]. Ren et al. [19] studied the generation and classification of alternative windows together with regression and proposed the Faster R-CNN network. In addition, there are two-stage target detection networks, such as CoupleNet [20], Cascade R-CNN [21], and other derived networks based on image pyramids and feature pyramids.

The YOLO series of algorithms is the most widely used one-stage target detection algorithm at present. Redmon et al. proposed the initial YOLO [13] model based on estimating the existence probability, types, and position deviation of objects in each detection area. On this basis, the batch normalization, high-resolution input, and full convolution operations are introduced to improve the target detection accuracy to achieve YOLOv2 [22]. YOLOv3 [14] uses Darknet53 as the backbone network and incorporates multiscale prediction improvements. YOLOv4 and YOLOv5 are mainly improved on data preprocessing and feature fusion [23]. YOLOv5 is designed to adapt to different target detection tasks by scaling the width and depth equally. YOLOv6 and YOLOv7 were derived from the YOLO series in 2022. Among them, YOLOv6 was proposed after designing a more efficient backbone and feature fusion network based on RepVGG style. YOLOv7 was proposed by the author of YOLOv4 referring to the ELAN structure extending to the extended efficient aggregation network, E-ELAN, which can improve the learning ability of the network without destroying the original gradient path. It is the best YOLO network model for overall performance. In addition, there are many versions of YOLO network derivation models [24], such as YOLOX [25] and YOLOR [13,26]. SSD is also a representative single-stage target detection network. It detects objects of different scales through feature maps of different layers and sizes to achieve the accurate detection of various targets. Based on the characteristics of the SSD algorithm [12], researchers have proposed different improvement methods, such as introducing feature layer context information and adding multilevel and feature pyramid structures.

According to the idea of visual tasks, Carion et al. [15] applied the Transformer structure for natural language processing to the process of computer vision tasks. Based on the Transformer structure, a target detection model, DETR, was proposed. The features were extracted using a convolutional neural network, and the position and category of the target were directly predicted by combining the Transformer codec network. In order to overcome the shortcomings of complex calculation, slow convergence speed, difficulty in training the network, and low accuracy of small target detection, Zhu et al. proposed a deformable attention module based on local sparsity with reference to deformable convolution [27], and then obtained Deformable DETR. Mehta S et al. [28] combined the advantages of CNN and Transformer to build a lightweight network architecture, MobileViT, to realize the application of Transformer in a lightweight target detection network. The above target detection algorithm is a universal target detection model for the whole scene, but there is still much room for optimization in the performance of the substation scene.

In this paper, YOLOv7, the target detection model in the YOLO series, is used as the basic model. We used the self-attention module to enhance feature extraction, the channel and spatial attention module to guide the fusion of high-level features and low-level features, and a light weight to improve the feature fusion network. This method improves the detection accuracy of the YOLOv7 model and reduces the network weight. The improved method can realize the detection of a power cabinet bounce lock and its status in the substation scenario.

The new contributions of this work are summarized as follows:

- The MobileViT module is used to improve the feature extraction ability of the YOLOv7 backbone network.
- A lightweight feature fusion network is designed based on the channel attention module and the spatial attention module.
- An improved GF-YOLOv7 network model is proposed for the recognition of substation lock state. The experimental results show that the recognition accuracy can be improved while the weight of the model can be reduced.

The remainder of this paper is organized as follows. The YOLOv7 model and the self-attention mechanism are described in detail in Section 2. In Section 3, we describe how the proposed algorithm enhances the feature extraction capability of the backbone network through the attention mechanism and designs the neck feature fusion network to improve the specificity of feature fusion. The proposed GF-YOLOv7 algorithm is compared with other methods for detection in Section 4. Conclusions and future work are discussed in Section 5.

2. Related Works

In this section, we briefly introduce important structures such as the YOLOv7 model and the self-attention mechanism, which are the basis of our research.

2.1. Principle of YOLOv7 Algorithm

The YOLO series target detection algorithm has been developed with many versions, and it plays an important role in the field of target detection. YOLOv7 is the YOLO target detection network released by YOLOv4's official team in July 2022. With the same size, YOLOv7 has higher detection accuracy than the most popular one, YOLOv5. Like YOLOv5, YOLOv7 is released in six basic versions to meet the requirements of different detection scenarios. The focus of the present study is detecting elastic locks of substation power cabinets, which has high requirements for a lightweight model and real-time detection. Thus, the lightest version, YOLOv7-tiny, was adopted as the prototype for the detection model. The structure of YOLOv7-tiny is shown in Figure 1.

The YOLOv7-tiny network model mainly includes four parts: input, backbone, neck, and prediction. The input uses Mosaic data enhancement, adaptive anchor frame calculation, and adaptive image scaling to preprocess the input image. These methods can improve the data quality of the input model. The backbone network is greatly improved compared to the previous-generation YOLO model. The focus downsampling structure is restored to a convolution layer with a step size of 2, and part of the semantic information is retained while downsampling. The extended efficient aggregation network, E-ELAN, is designed according to the ELAN [29] structure to improve the learning ability of the network without destroying the original gradient path. During the two downsampling processes from P3 to P4 and P4 to P5, E-ELAN handles feature extraction, and the downsampling operation is completed by maximum pooling, so as to further reduce the calculation parameters and calculation amount while ensuring feature extraction. The neck continues to use the path aggregation network structure of the FPN + PAN structure, but replaces the E-ELAN layer with the CSP module. Prediction sets the grid of 1/8, 1/16, and 1/32 of the input image according to the size of the detected object. Each grid contains three prediction boxes, and each prediction box contains the classification, location, and confidence information of the target. Finally, the redundant prediction boxes are eliminated by NMS (non-maximum

suppression), and the information of the prediction box with the highest confidence is retained, so as to complete the target detection process.

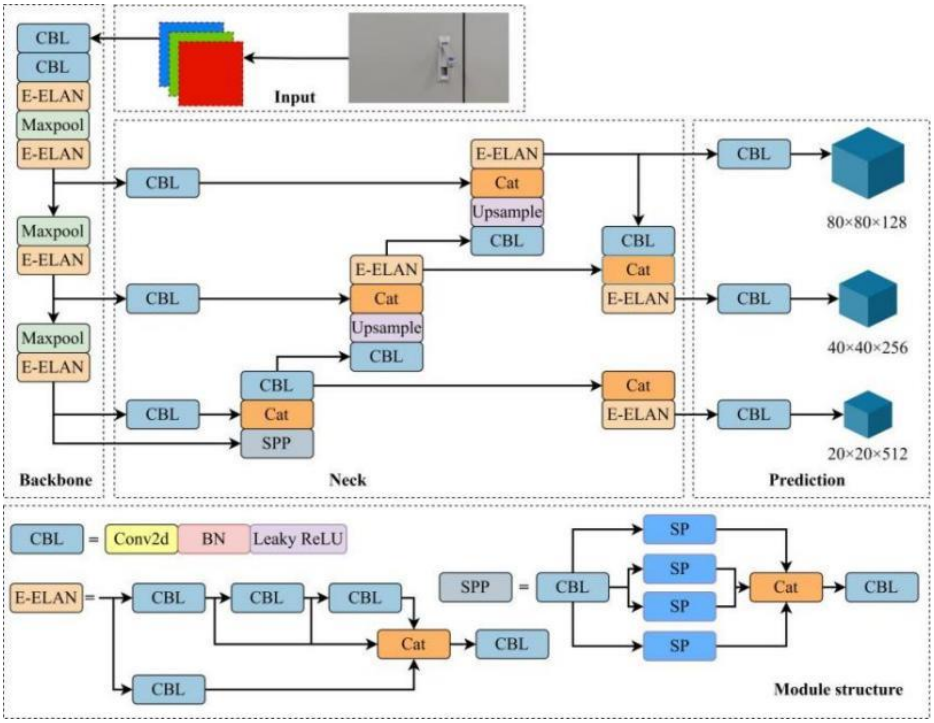


Figure 1. Structure of YOLOv7-tiny model.

In addition, YOLOv7-tiny is also optimized for model training. On the one hand, model re-parameterization design refers to RepConv to compress the size of the model while ensuring the accuracy of the model, and to achieve complex model training and inference. On the other hand, using the idea of deep supervision, an additional auxiliary header is added to the middle layer of the network to detect the learned information, and a new soft label generation method is used to train the model, which can improve the detection ability of the network.

The existing YOLOv7-tiny algorithm is designed for the target detection task across all scenarios. In order to make the network more suitable for the target detection of power cabinets locks in a substation, targeted improvement of YOLOv7-tiny is needed.

2.2. Self-Attention Mechanism

Convolutional neural networks use natural inductive bias advantages to learn visual representations and establish local dependencies on spatial information domains, but they also lack the ability to learn global representations. Visual Transformer (ViT), which is based on the self-attention mechanism, has the ability to capture global receptive fields of input feature maps and can build global dependencies on spatial dimensions to learn global visual representation information. ViT architectures are usually intensive and difficult to train due to a lack of spatial generalization bias. On this basis, Mehta S et al. combined the advantages of CNN and Transformer to build a lightweight network architecture, MobileViT, as shown in Figure 2.

In the MobileViT architecture, MobileViT Block (MVBlock) is the core part of integrating Transformer into the image feature extraction network. The specific process of extracting global features is as follows:

(1) The local spatial information of the input tensor $X \in R^{H \times W \times C}$ is learned by using a general convolution;

(2) The output feature map adjustment channel of step (1) is projected into a high-dimensional space through point convolution, and the tensor $X_L \in R^{H \times W \times d}$ is obtained by learning global feature transformation through Transformer after expanding the tensor $X_G \in R^{H \times W \times d}$;
(3) Refold tensor $X_G \in R^{H \times W \times d}$ and obtain tensor $X_F \in R^{H \times W \times C}$ through point convolution;
(4) After splicing the result of step (1) with tensor $X_F \in R^{H \times W \times C}$, the local features and global features are fused by convolution.

Through the relay extraction of local information and global information by the convolution module and the transformer module, MVBlock not only has the spatial induction bias feature of convolution, but also can model the entire feature layer to realize the simultaneous perception of local information and global information of the feature map, so it can achieve better performance with fewer channels and a shallower network.

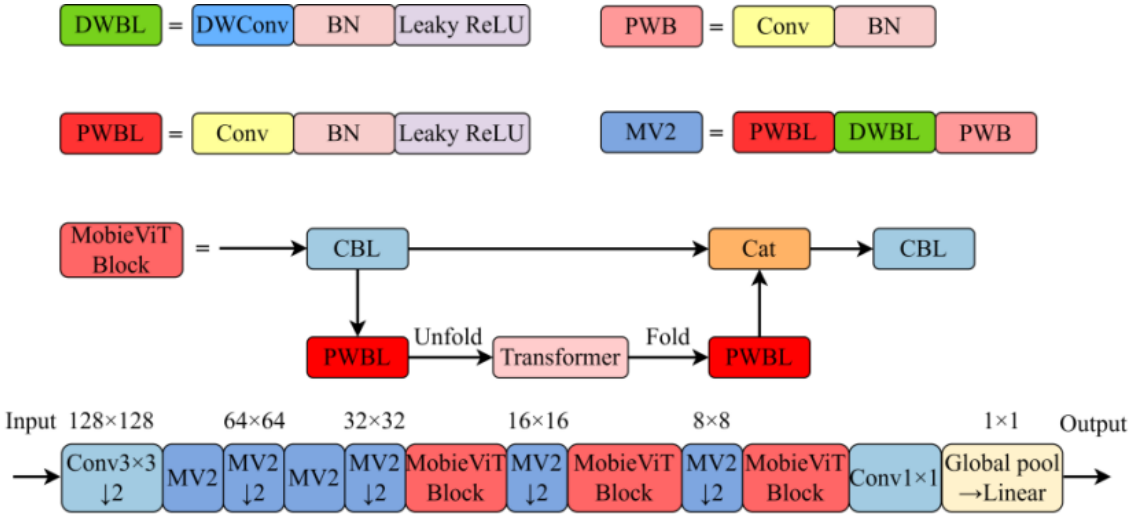


Figure 2. The architecture of MobileViT.

2.3. Channel Attention Mechanism and Spatial Attention Mechanism

The feature layer extracted from the backbone network contains rich semantic information, but the semantic information contained in the feature map at different levels is different, and the contribution to target detection is also different. YOLOv7-tiny's neck structure integrates all levels, which can improve the target detection ability, but not all levels of information have the same contribution to target detection, and redundant information may even mislead the network to use effective information.

The attention mechanism can adjust the weight of the fused information, enhance the attention of the neural network to useful information, and inhibit the attention to invalid information. In order to enable the detector to divide attention when detecting different targets, the attention mechanism thus improves the perception of useful information.

As shown in Figure 3, the CBAM attention mechanism can perceive both channel attention (CA) and spatial attention (SA) [30]. The CA module is similar to SENet (Squeeze-and-Excitation Network), which adds the feature extraction method of maxpool on the basis of SENet and uses the channel relationship between features to generate the channel attention map. The SA module utilizes the spatial relationship between features, parallels the average pooling and maximum pooling operations along the channel axis, and obtains the attention map through a convolution layer after connection.

In this paper, the CBAM attention mechanism is applied to multi-scale feature fusion, and the fused feature map is adjusted.

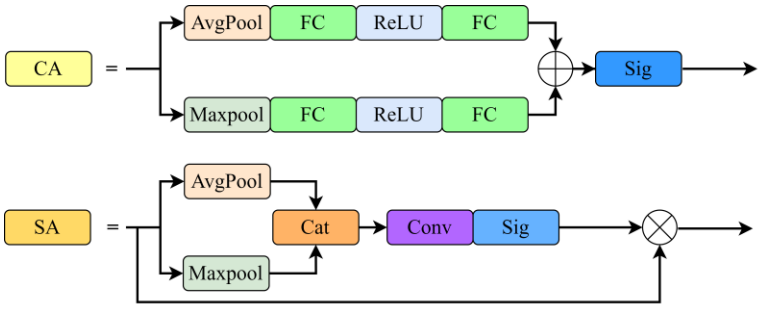


Figure 3. Structure of CBAM.

3. Proposed Methods

In this paper, we use the attention mechanism to enhance the feature extraction ability of the backbone network and design the neck feature fusion network to improve the specificity of feature fusion. The end of this section shows the overall architecture of the improved GF-YOLOv7 network model.

3.1. Improvement of Feature Extraction Network

While improving the feature extraction ability of the backbone network, the self-attention mechanism module will also increase the amount of network parameters and computation, resulting in a decline in network reasoning speed and difficulties in engineering applications.

After many tests and comparisons, when MVBlock is added after the third and fourth downsampling and feature extraction, the feature extraction effect and reasoning speed of the network reach a balance. The GAE-ELAN (Global Attention Extraction–Efficient Long-Range Attention Network) structure of the improved feature extraction network is shown in Table 1. CBL (Convolution Batch Normal Leaky ReLU) is a common layer stack structure, common in convolutional neural networks and some deep learning models. Maxpool (max pooling) is a pooling operation. Pooling layers are often inserted between successive convolutional layers to reduce the space size of their input while preserving important information.

Table 1. Overall architecture of the feature extraction network GAE-ELAN.

Input	Network Unit	Channel Number	Step
$1^2 \times 3$	CBL	32	2
$1/2^2 \times 32$	CBL	64	2
$1/4^2 \times 64$	E-ELAN	64	1
$1/4^2 \times 64$	Maxpool		2
$1/8^2 \times 64$	E-ELAN	64	1
$1/8^2 \times 64$	MVBlock	64	1
$1/8^2 \times 64$	Maxpool		2
$1/16^2 \times 64$	E-ELAN	128	1
$1/16^2 \times 128$	MVBlock	128	1
$1/32^2 \times 128$	Maxpool		2
$1/32^2 \times 128$	E-ELAN	128	1

3.2. Improvement of Feature Fusion Network

The channel attention module (CA) and the spatial attention module (SA) are embedded into the feature fusion network of the neck according to the needs, and the neck is lightweight to make it more focused on feature fusion. The feature fusion capability of the neck is preserved while reducing the computational overhead. The obtained feature fusion network is named FF-FPN (Feature Fusion–Feature Pyramid Networks), and its network structure is shown in Figure 4.

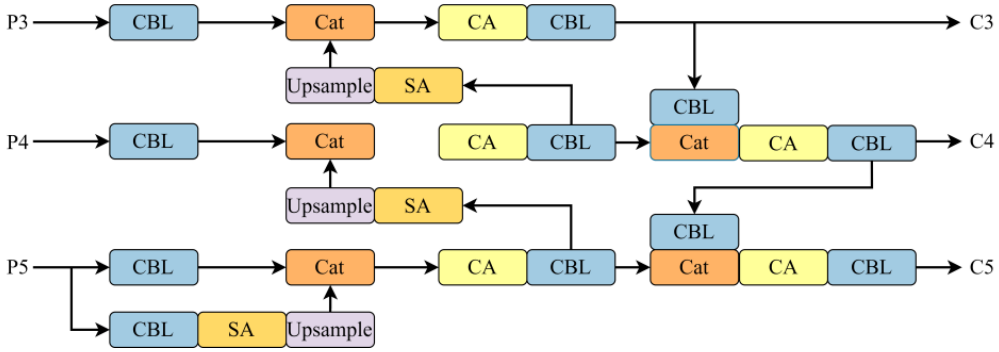


Figure 4. Structure of FF-FPN.

The feature fusion process of FF-FPN is as follows:

(1) The P5 feature layer extracted from the backbone network undergoes a convolution downsampling to obtain the feature layer P6 with a higher level of semantic information.

(2) The feature layer with higher semantic information guides the information fusion of the next feature layer with the help of the attention mechanism. The specific method is to obtain the spatial attention weight of P6, P5, and P4 feature layers through the SA module once, and then splice them with P5, P4, and P3 layers after upsampling. The spliced feature layers are perceived by the channel attention weight through the CA module once, and the feature channels are fused using convolution to obtain the preliminary fused feature maps P′5, P′4, and P′3.

(3) Integrate the low-level feature map to the high level. The bottom feature maps P′4 and P′3 are spliced with the higher-level P′5 and P′4 through downsampling, the channel attention weight is obtained through a CA module, and then the feature channels are fused through convolution to obtain the further fused feature layers P″5, P″4, and P″3.

(4) According to the needs of the detection task, repeat steps (2) and (3) to obtain the final fused feature layers C5, C4, and C3.

3.3. GF-YOLOv7 Network Model

To detect the bouncing lock target of a substation power cabinet based on the YOLOv7-tiny algorithm, the backbone network and neck are replaced by the improved feature backbone network and feature fusion network. The improved GF-YOLOv7 (Global Attention Extraction and Feature Fusion–YOLOv7) network is obtained and the overall structure is shown in Figure 5.

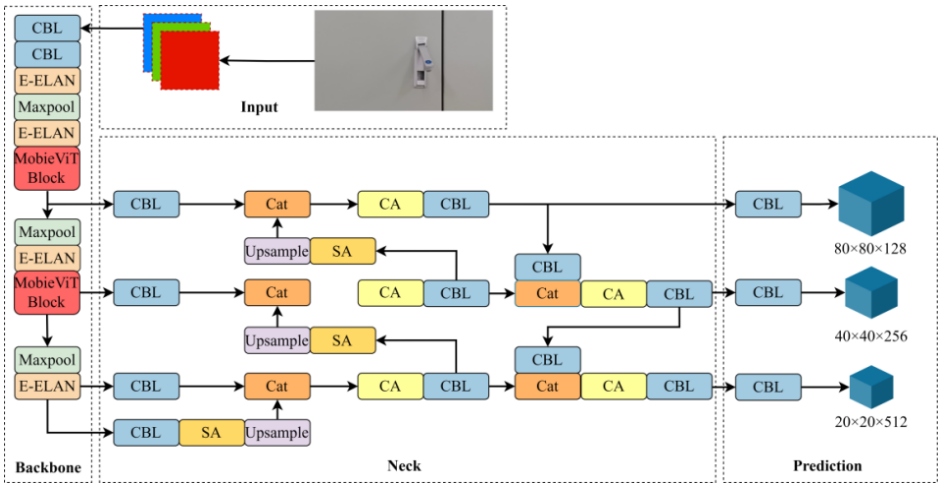


Figure 5. Overall structure of GF-YOLOv7.

4. Experimental Results

In this section, we firstly describe the experiment's implementation details. Then, we introduce the evaluation criterion and datasets for evaluation. Subsequently, we study the proposed GF-YOLOv7 and compare it with other methods. Finally, we discuss the effectiveness and limitations of our proposed method.

4.1. Implementation Details

The target detection network training and verification dataset came from the video shot in the substation scene, and the video was extracted frame by frame to obtain 7756 pictures with a resolution of 1920 × 1080 dpi, including four kinds of open and closed bouncy locks, for a total of eight datasets, as shown in Figure 6. According to the ratio of 8:1:1, the dataset was divided into a training set, verification set, and test set, used for the training of the algorithm network and the testing of the algorithm's performance.

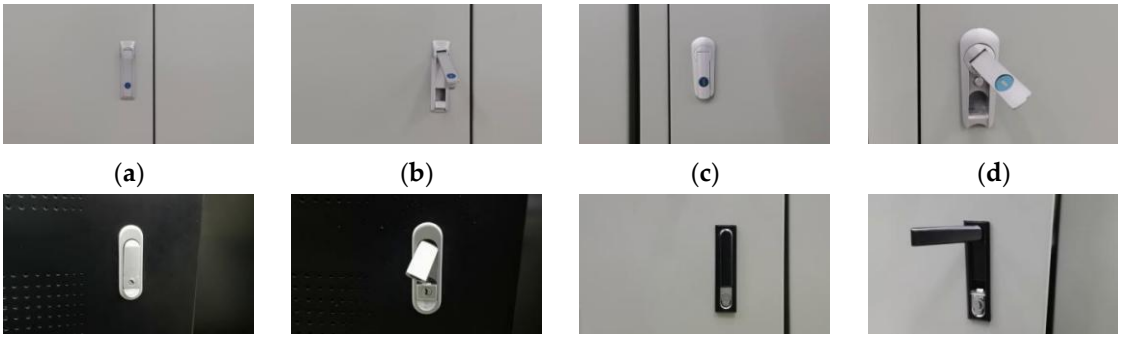


Figure 6. The opening and closing states of four kinds of substation power cabinet spring locks. (**a**) THA0; (**b**) THA1; (**c**) THB0; (**d**) THB1; (**e**) THC0; (**f**) THC1; (**g**) THD0; (**h**) THD1.

In the experiment, the Make Sense annotation tool was used to select the power cabinet bounce lock with a rectangular box and generate the label file in xml format, as shown in Figure 7.

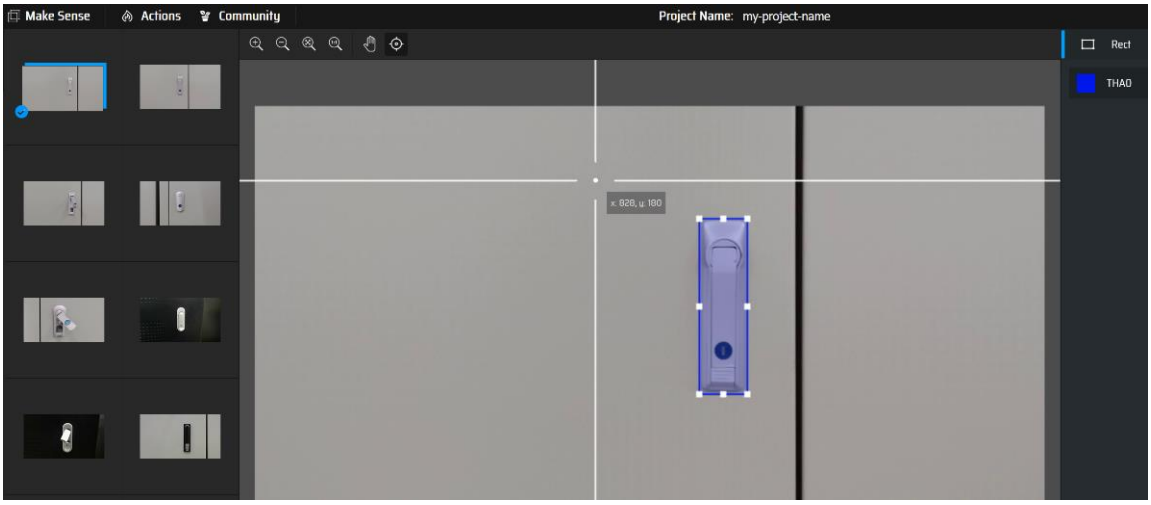


Figure 7. Label fabrication.

The model training and testing environment was the Ubuntu 20.04 operating system, and the calculation program adopted Python 3.8 language combined with Python 1.8.1 deep learning framework. During the training process, a NVIDIA GeForce RTX3070 graphics card was used to train the two networks. The super parameter setting in the training stage

was as follows: the initial learning rate was 0.0031, the attenuation coefficient was 0.12, the momentum was 0.833, the batch size was 16, and the training number was 300.

4.2. Evaluation Indicators

In target detection, the average accuracy AP of each type of target and the average accuracy mAP of all targets are used to evaluate the detection effect and performance of the model. AP is the area under the recall and precision curves. The calculation formula is as follows:

$$Pr = \frac{TP}{TP + FP} \tag{1}$$

$$Re = \frac{TP}{TP + FN} \tag{2}$$

$$AP = \frac{1}{101} \sum_{i=0}^{100} Pr\left(Re = \frac{i}{100}\right) \tag{3}$$

$$mAP = \sum_{i=1}^{N} \frac{AP_i}{N} \tag{4}$$

In Formulas (1)–(4), Pr is the detection accuracy, Re is the recall rate of the model, TP is the number of accurate predictions, FP is the number of false detections, FN is the number of missed detections, and N is the target category.

The detection speed is expressed in FPS (frames per second), or the number of pictures that the model can process per second. The complexity of the model is measured by the weight of the model, the number of parameters, and the floating point number.

4.3. Test Results and Analysis of Power Cabinet Spring Lock

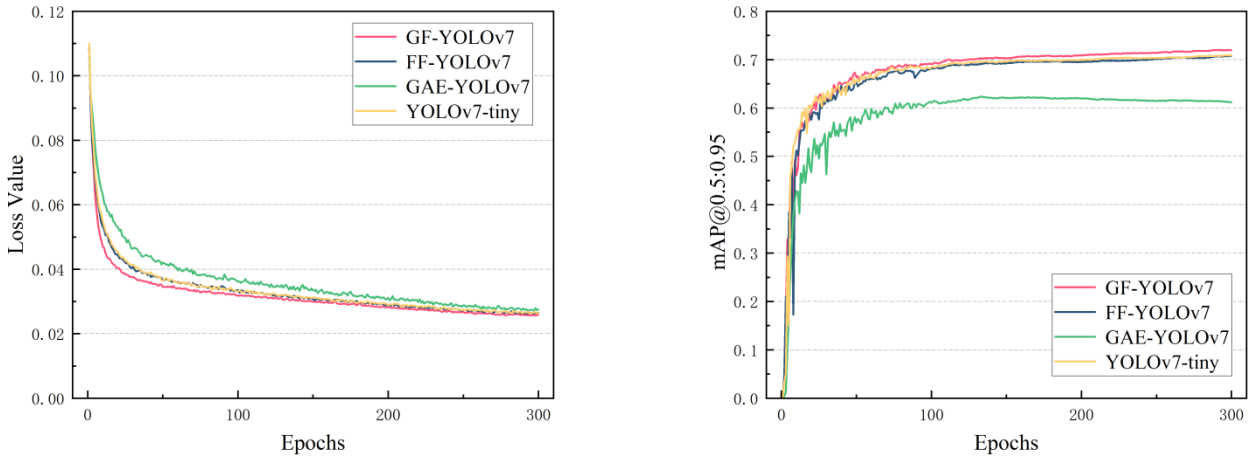
In order to verify the effectiveness of the model improvement and study the impact of each improvement measure on the detection effect, a comparative experiment was set up to study the improved network. First, we added MVBlock with the same parameters as GF-YOLOv7 to the backbone network of the YOLOv7-tiny network to obtain the GAE-YOLOv7 network. FF-YOLOv7 was obtained by replacing the neck structure of YOLOv7-tiny with FF-FPN. Then, we used the dataset mentioned in the previous section to train GF-YOLOv7, FF-YOLOv7, GAE-YOLOv7, and the original YOLOv7-tiny network. Table 2 shows the network complexity comparison of the four models.

Table 2. Algorithmic model complexity comparison.

Models	Backbone Network	Neck	Layer	Parameter Quantity	Model Size/mb
YOLOv7-tiny	E-ELAN	FPN+PAN	200	6,025,525	12.3
FF-YOLOv7	E-ELAN	FF-FPN	184	4,166,174	8.5
GAE-YOLOv7	GAE-ELAN	FPN+PAN	338	9,296,693	18.9
GF-YOLOv7	GAE-ELAN	FF-FPN	322	4,929,915	9.1

Table 2 shows that after adding MVBlock to the backbone network, the number of layers and parameters of the network model increased, which also led to an increase in the weight file size. After replacing the feature fusion part of the network with FF-FPN, the network weight file size decreased. The weight file of GF-YOLOv7 network after re-parameterization was only 9.1 M, which is conducive to the engineering application of the network.

The changes in loss function and mAP@0.5:0.95 of the four network models were compared and analyzed, and the corresponding change curves are shown in Figure 8.



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

Figure 8. Comparison of test results. (**a**) Change curve of loss function; (**b**) Change curve of mAP@0.5:0.95.

It is obvious from Figure 8a that the loss function value of the GF-YOLOv7 model decreased faster and the final loss value was the lowest. Figure 8b shows the average accuracy of the four models at the confidence level of 0.5~0.95. Except for the GAE-YOLOv7 network, the loss functions of the other three models increased rapidly at the beginning of training, with small overall fluctuations and good convergence. The average accuracy of GF-YOLOv7 is better than other networks.

In conclusion, compared with YOLOv7-tiny, the FF-YOLOv7 model, and the GAE-YOLOv7 model, the GF-YOLOv7 model has better detection performance and recognition effect for substation bounce locks. Table 3 compares the training results of the four networks after 300 rounds of training.

Table 3. Comparison of model evaluation indicators.

Models	Precision (%)	Recall (%)	mAP@0.5 (%)	mAP@[0.5:0.95] (%)
YOLOv7-tiny	0.9916	0.9948	0.9931	0.7099
FF-YOLOv7	0.9917	0.9961	0.9939	0.7081
GAE-YOLOv7	0.8318	0.9817	0.8741	0.6119
GF-YOLOv7	0.9914	0.9960	0.9947	0.7191

The training results show that the GAE-YOLOv7 network with the self-attention mechanism added directly into YOLOv7-tiny has better performance than FF-YOLOv7, although the loss function is smaller than other networks, but the actual performance does not increase or decrease. Analyzing the training process, it is found that GAE-YOLOv7 has poor detection results for very small targets such as floating targets, which affects the overall performance of the network. When the feature fusion network is adjusted to FF-FPN, the effective fusion of information makes the performance of the network slightly better than before. In general, we can see that adding a self-attention mechanism to the network, perceiving global features, and adding the feature fusion network of attention can ensure the detection ability of small targets, and the performance of the target detection network is improved significantly.

The experiment used four networks to detect 775 pictures of test data with a total of 901 bounce lock targets. Figure 9 shows the large target bounce lock with a simple background based on YOLOv7-tiny, FF-YOLOv7, GAE-YOLOv7, and GF-YOLOv7 network models.

Figure 10 shows the results of mixed bounce lock detection with a complex background based on YOLOv7-tiny, FF-YOLOv7, GAE-YOLOv7, and GF-YOLOv7 network models.

The test results of Figure 9 show that for single bounce lock detection with a simple background and large target, YOLOv7-tiny, FF-YOLOv7, and GF-YOLOv7 networks have better recognition performance, and the confidence of GF-YOLOv7 recognition is slightly higher than other networks. The test results of Figure 10 show that for mixed bounce lock detection with a complex background and a small target, the recognition accuracy and confidence of GF-YOLOv7 network are significantly higher than other networks. The results show that this improved method can better complete the recognition and classification tasks.

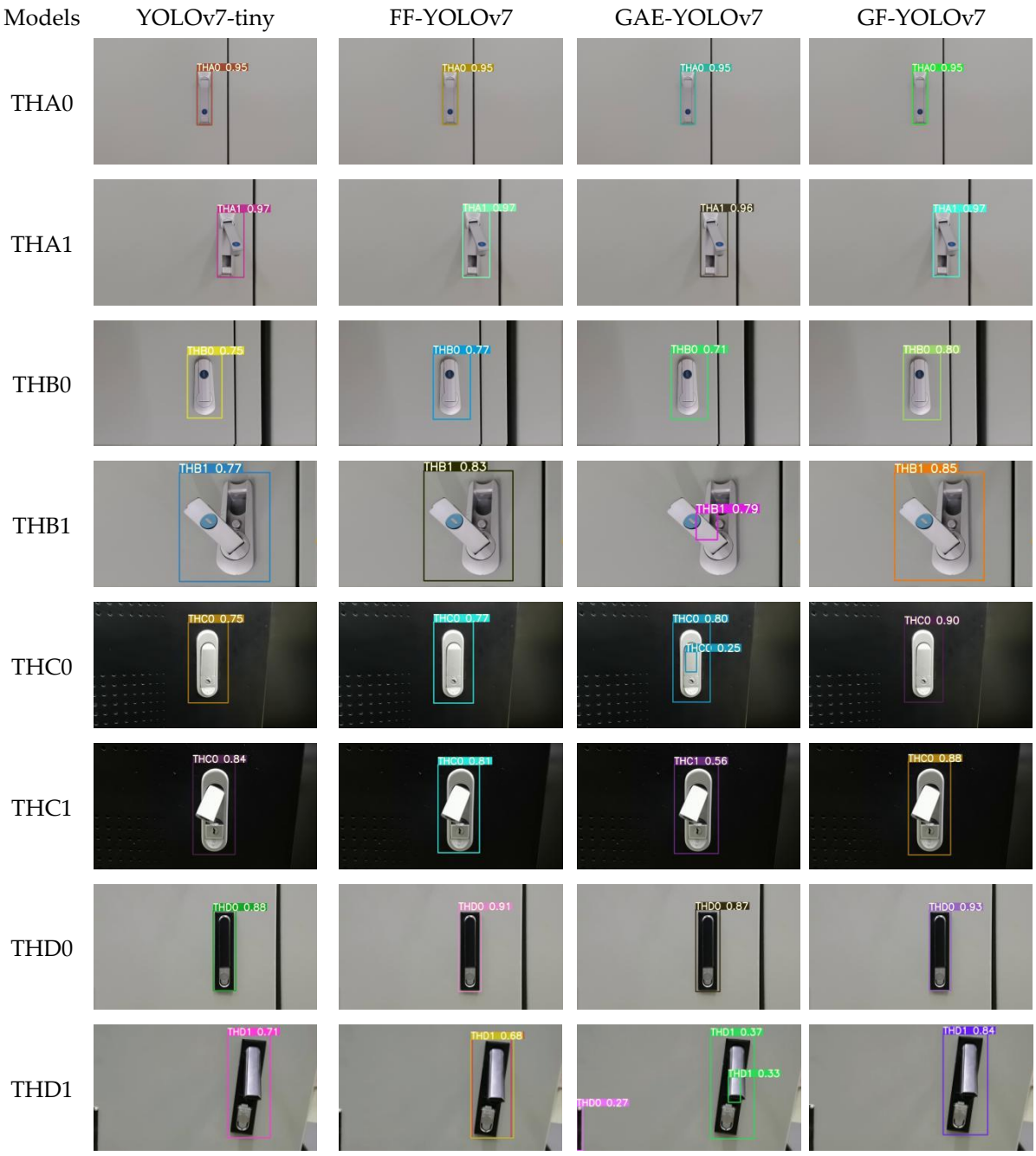


Figure 9. Comparison of single bounce lock detection with a simple background and large target.

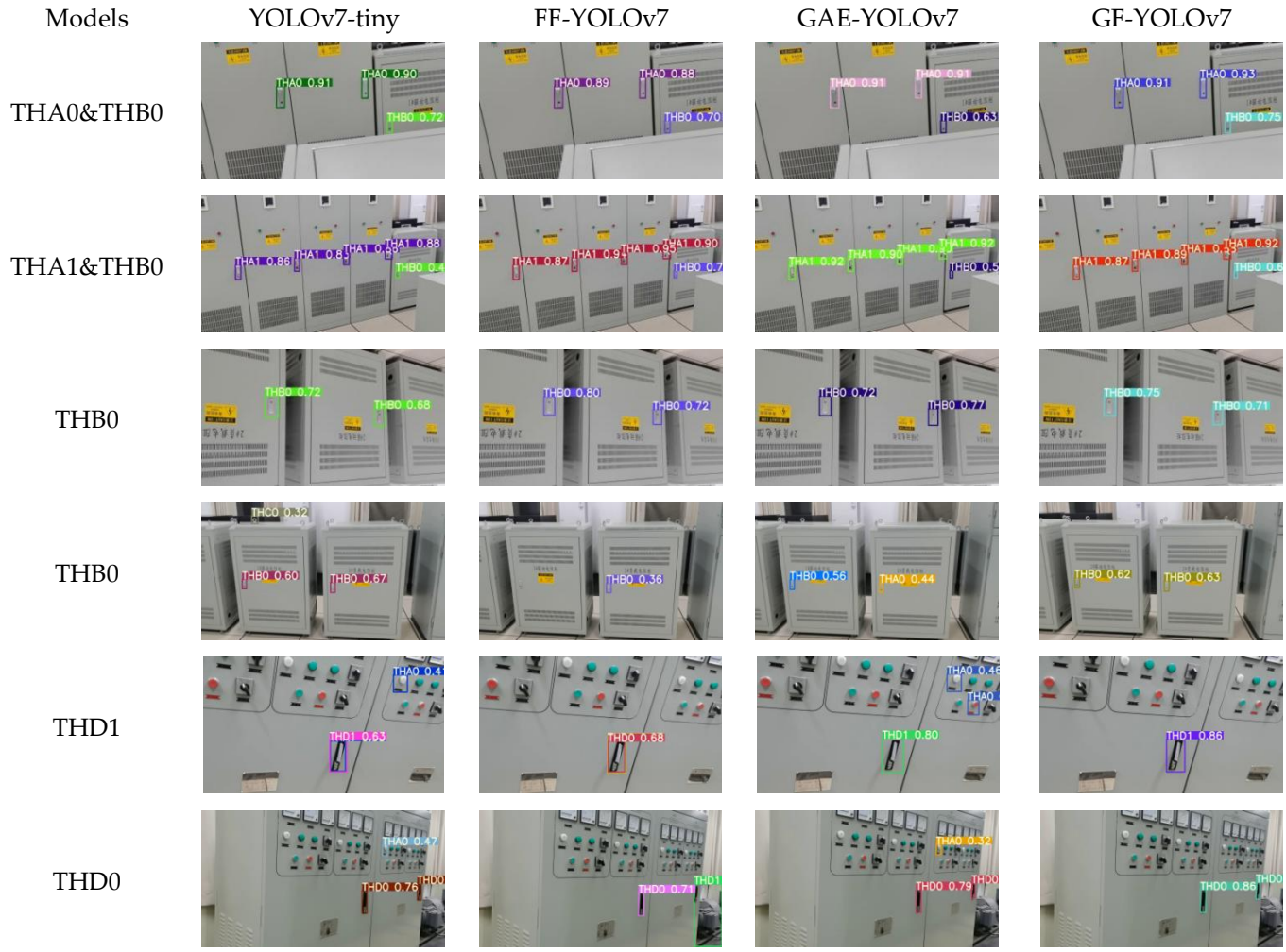


Figure 10. Comparison of mixed bounce lock detection with a complex background.

Table 4 shows the detection result statistics of 775 pictures of test data by four networks.

Table 4. Comparison of model detection accuracy and speed.

Models	Misidentification	Accuracy (%)	FPS	Inference Time (s)
YOLOv7-tiny	63	93.0	33.9	22.9
FF-YOLOv7	32	97.6	32.4	23.9
GAE-YOLOv7	143	81.0	33.8	22.9
GF-YOLOv7	19	97.8	32.1	24.1

Four models were used to detect 775 test pictures. In terms of detection accuracy, the error rate of GAE-YOLOv7 is significantly higher than other networks. The detection accuracy of YOLOv7-tiny and FF-YOLOv7 are approximately the same, and that of GF-YOLOv7 is higher than other networks. In terms of detection speed, the four models can meet the needs of real-time detection, and the GF-YOLOv7 network is slightly slower than the other networks. It is obvious that the accuracy of the GF-YOLOv7 network is 97.8% in the test set, the detection speed is 32.1 FPS, and the model weight file size is 9.1 M. It is most suitable for the real-time detection of bounce locks in substations from images when target detection is completed.

5. Conclusions and Future Works

In this study, we proposed a target detection algorithm based on GF-YOLOv7 for the detection task of substation bounce locks. As a model for target detection algorithms, YOLOv7-tiny incorporates a self-attention mechanism into the task of target detection through the MobileViT module. This mechanism, utilizing the Transformer structure's capacity to capture global information, allows for the comprehensive extraction of both local and global data. By merging the separated CBAM attention mechanism with the feature fusion network, we designed the FF-FPN feature fusion network, ultimately leading to the construction of the GF-YOLOv7 target detection network. Our experimental validation revealed significant findings. In terms of detection accuracy, GAE-YOLOv7 exhibits a notably higher error rate than the other networks. Both YOLOv7-tiny and FF-YOLOv7 demonstrate roughly comparable detection accuracies, while the GF-YOLOv7 surpasses other networks with a detection accuracy of 97.8%. Regarding detection speed, all four models fulfill the criteria for real-time detection. The GF-YOLOv7 operates at a speed of 32.1 frames per second and boasts a model weight file size of 9.1 megabytes. These results indicate that the GF-YOLOv7 model meets the demands for detection accuracy and real-time performance. The model's lightweight file structure presents another advantage, rendering it more suitable for engineering applications.

In the future, we plan to improve the detection accuracy and speed of the algorithm on small targets. The weight of the model further reduced to facilitate better deployment of identification tasks on the edge computing platform.

Author Contributions: Conceptualization, X.Z.; Project administration, L.L.; Writing—original draft, Y.W.; Methodology, P.Z.; Validation L.W.; Table, Z.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China (grant no. 41771487).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy of the subjects involved in study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhu, M.; Qin, Q.; Huang, C.; Zhang, W.; Liang, Z.; Chen, J. A Detection Method of Unsafe Behavior in Substation Based on Deep Learning. In Proceedings of the 3rd International Conference on Information Technologies and Electrical Engineering, Changde, China, 3–5 December 2020; pp. 499–502.
2. Gong, Q.; Li, J.; Luo, Y.; Gu, Q. State Detection Method of Secondary Equipment in Smart Substation Based on Deep Belief Network and Trend Prediction. In Proceedings of the 2019 IEEE Sustainable Power and Energy Conference (iSPEC), Beijing, China, 21–23 November 2019; pp. 2369–2373.
3. Fu, C.-Z.; Si, W.-R.; Huang, H.; Chen, L.; Gao, Q.-J.; Shi, C.-B.; Wang, C. Research on a Detection and Recognition Algorithm for High-Voltage Switch Cabinet Based on Deep Learning with an Improved YOLOv2 Network. In Proceedings of the 2018 11th International Conference on Intelligent Computation Technology and Automation (ICICTA), Changsha, China, 22–23 September 2018; pp. 346–350.
4. Song, W.; Liu, X.; Zhao, J.; Wang, M.; Liu, Y. Research on the Intelligent Identification Method of the Substation Equipment Faults Based on Deep Learning. In Proceedings of the 2020 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS), Shenyang, China, 28–30 July 2020; pp. 888–891.
5. Yilin, J.; Jian, S. Substation Equipment Fault Identification Based on Infrared Image Analysis. In Proceedings of the Journal of Physics: Conference Series, Moscow, Russia, 20–21 October 2020; p. 012004.
6. Li, Y.; Xu, Y.; Xu, M.; Wang, S.; Xie, Z.; Li, Z.; Jiang, X. Automatic infrared image recognition method for substation equipment based on a deep self-attention network and multi-factor similarity calculation. *Glob. Energy Interconnect.* **2022**, *5*, 397–408. [CrossRef]

7. Zheng, H.; Sun, Y.; Liu, X.; Djike, C.L.T.; Li, J.; Liu, Y.; Ma, J.; Xu, K.; Zhang, C. Infrared Image Detection of Substation Insulators Using an Improved Fusion Single Shot Multibox Detector. *IEEE Trans. Power Deliv.* **2020**, *36*, 3351–3359. [CrossRef]
8. Ciric, R.; Milkov, M. Application of Thermal Imaging in Assesment of Equipment in Power Plants. *Monit. Expert. Saf. Eng.* **2014**, *4*, 1–8.
9. Wang, L.; Kou, Q.; Zeng, Q.; Ji, Z.; Zhou, L.; Zhou, S. Substation switching device identification method based on deep learning. In Proceedings of the 2022 4th International Conference on Data-Driven Optimization of Complex Systems (DOCS), Chengdu, China, 28–30 October 2022; pp. 1–6.
10. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
11. 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, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
12. 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 Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
13. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 779–788.
14. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
15. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 213–229.
16. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-Captured Scenarios. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 2778–2788.
17. Yue, X.; Wang, Q.; He, L.; Li, Y.; Tang, D. Research on Tiny Target Detection Technology of Fabric Defects Based on Improved YOLO. *Appl. Sci.* **2022**, *12*, 6823. [CrossRef]
18. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
19. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 1–9. [CrossRef] [PubMed]
20. Zhu, Y.; Zhao, C.; Wang, J.; Zhao, X.; Wu, Y.; Lu, H. CoupleNet: Coupling Global Structure with Local Parts for Object Detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4126–4134.
21. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving Into High Quality Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162.
22. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
23. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
24. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv* **2022**, arXiv:2207.02696.
25. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOX: Exceeding YOLO Series in 2021. *arXiv* **2021**, arXiv:2107.08430.
26. Wang, C.-Y.; Yeh, I.-H.; Liao, H.-Y.M. You Only Learn One Representation: Unified Network for Multiple Tasks. *arXiv* **2021**, arXiv:2105.04206.
27. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable Convolutional Networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 764–773.
28. Mehta, S.; Rastegari, M. MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer. *arXiv* **2021**, arXiv:2110.02178.
29. Zhang, X.; Zeng, H.; Guo, S.; Zhang, L. Efficient Long-Range Attention Network for Image Super-resolution. In Proceedings of the Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, 23–27 October 2022; pp. 649–667.
30. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.