Angle Steel Tower Bolt Defect Detection Based on Transformer

Jinfeng Zhang¹, Yuanwei Hu², and Shujun Ji²
¹State Grid Anhui Electric Power Co., Ltd., Hefei, Anhui 230022, China
²Department of Applied Mathematics, Wenzhou-Kean University, Wenzhou, Zhejiang 325000, China
*Email: ahu0086@163.com

Abstract. The bolts of the angle steel tower will rust, loose, and fall off under natural conditions. Traditional manual bolt defect detection is inefficient and dangerous. This paper proposes CViT-FRCNN based on ViT-FRCNN, which uses a convolutional neural network as the backbone model and the output features are input to the Transformer encoder. Compared with the direct patch embedding of ViT-FRCNN, this can improve the richness of input features and detection accuracy. A series of experiments show that our proposed model achieves the best performance in angle steel tower bolt defect detection and meets the needs of power inspection scenarios.

1. Introduction
The bolts of the angle steel tower play an important role in the normal operation of the power system. However, under the influence of various natural factors, the bolts appear rusty, loose and fall off [1]. Therefore, it is an indispensable step to conduct a safety inspection on the bolt condition of the angle steel tower. However, the safety check of bolts mainly relies on manual operation, which is time-consuming and labor-intensive, and inefficient. In recent years, some drones and climbing robots with cameras have been applied to power inspections [2, 3].

In order to overcome the limitations of manual inspection, along with the development of deep learning, automated inspection technology has made great progress. Zhao et al. [4] used multiple Convolutional Neural Network (CNN) modules to extract image features for power transmission line insulation detection. Tao et al. [5] proposed the detection of insulated lines based on the CNN cascade structure. However, due to the small targets of angle steel tower bolts and the complex and changeable positions of bolts, there is very little research on bolt defect detection. In addition, changes in lighting and shooting posture also increase the difficulty of detection. Object detection algorithms based on deep learning have been widely used in the field of computer vision, and are mainly divided into two categories: single-stage and double-stage. The main representative of the two-stage is the Regional Convolutional Neural Network (R-CNN) [6] series, while the single-stage is mainly the single-shot detector (SSD) [7] and YOLO (You only look once) [8]. Recently, the successful application of Transformer in image classification [9] has led to new developments in target detection algorithms.

Transformer [10] was originally applied in the field of natural language processing and has achieved great success in this field. Transformer based on self-attention mechanism has more advantages than CNN and recurrent neural network (RNN). Detection Transformer (DETR) is the first model that applies to target detection, and has obtained competitive results on the COCO dataset [11]. However, DETR has poor convergence and poor detection of small targets [12]. Vision Transformer
(ViT) [9] is the first pure Transformer model to achieve advanced results on large-scale data sets. ViT directly cuts the input picture into 16×16 small blocks, and then inputs it to the Transformer's encoder through linear projection, and finally predicts the category via the Multi-layer perceptron (MLP). The successful application of ViT has attracted much follow-up attention. ViT-FRCNN [13] replaces the backbone network of Faster R-CNN with ViT, and finally predicts the result via a region proposal network (RPN).

This paper proposes a Transformer-based target detection algorithm to detect defects in angle steel tower bolts. We replace the block embedding in the input part of ViT-FRCNN with a CNN backbone network, and use the features extracted by CNN to increase the complexity of the input information, named CViT-FRCNN. We also collected pictures of angle steel tower bolts, cleaned and labeled them, and used data enhancement to make up for the shortcomings of insufficient data. Finally, a series of experiments proved the effectiveness of the method.

2. Target Detection Model Based on Transformer

2.1. Model Structure

![Figure 1. Structure of CViT-FRCNN model.](image)

The overall picture of the CViT-FRCNN model is shown in figure 1. The model is mainly composed of three parts: CNN backbone network, Transformer encoder and RPN detection network. In order to enable the Transformer encoder of CViT-FRCNN to extract more features, the direct block operation before the Transformer encoder is replaced with a CNN network, so that richer spatial information can be obtained to improve the accuracy of detection. The features output by CNN are divided into feature blocks with the same length and width without overlapping, and then these feature blocks are flattened and combined into a two-dimensional patch embedding. The label CLS representing the category is added to the block embedding for RPN network classification. In addition, because the Transformer lacks position information, one-dimensional position embedding is added to the block embedding, and then fed to the Transformer encoder to learn global features. The features output by the encoder are recombined into a new feature map according to the position code, and finally sent to the RPN model to predict the category and coordinates. In addition, a residual block is added between the Transformer encoder and the detection network, which helps to improve information flow and detection performance. The default setting of CNN backbone network we choose is ResNet-18, and the number of Transformer encoders is set to 8.

2.2. Transformer Encoder Structure

As shown in figure 2, the Transformer's encoder is mainly composed of multi-head self-attention (MSA), multi-layer perceptron, residual connection (Residual connection) [14] and layer normalization [15]. The core self-attention mechanism is expressed as

$$\text{Attention}(Q, K, V) = \text{softmax} \left( \frac{QK^T}{\sqrt{d_k}} \right) V.$$  (1)
Among them $Q$ is the query matrix, $K$ is the key matrix, $V$ is the value matrix, and $d_k$ is the dimension of $V$. $QK^T$ calculates the attention scores between different input matrices. The scaling factor $1/\sqrt{d_k}$ is mainly used to improve stability, and then the function $softmax$ converts the attention scores into probabilities. Finally, multiply it by $V$ to get the weight matrix.

In order to improve the feature extraction ability of self-attention, multiple self-attentions are concatenated into multi-headed attention, which can be written as

$$MSA(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V).$$

(2)

Among them $W$ is the weight matrix, and the number of multi-heads is set to 6 by default.

In addition to the self-attention mechanism and multi-head self-attention, layer normalization is mainly used to stabilize training and accelerate convergence, and residual connection is to improve information flow. The multilayer perceptron consists of two linear layers and a Gaussian error linear unit (GELU) activation function. Finally, by repeatedly stacking encoders, the global context dependence and the aggregation feature information are continuously extracted.

![Figure 2. Structure of Transformer encoder.](image)

3. Experiment

3.1. Data Set and Parameter Settings

In this paper, the angle steel tower bolts were photographed by both drones and manually and a total of 5892 color pictures were obtained. These images were manually inspected, and some fuzzy, residual, and indistinguishable low-quality images were removed, and finally a bolt data set of 5680 sheets was obtained. Bolt images are divided into two categories, one is normal bolts, and the other is faulty bolts. The data set is randomly shuffled, and 80% is selected as the training set and 20% as the test set. The pictures of angle steel tower bolts in part of the data set are shown in figure 3.

![Figure 3. Part of the bolt pictures.](image)

Due to the limited quantity of samples, the generalization ability of the model cannot be guaranteed, so data enhancement is used. The main methods are random cropping, random horizontal or vertical flipping and adding random noise. These data enhancement techniques are used to improve the accuracy and robustness of the model and overcome the shortcomings of insufficient data.
The experiment uses the Adam optimizer, other parameters remain the default, the batch size is 64. In addition, a learning rate decay strategy is also used to make the initial learning rate decay 10 times every 50 epochs, for a total of 200 epochs. In order to ensure the fairness and reasonableness of the comparison, let each model be run 5 times randomly, and the final results are averaged. The experimental equipment is Intel i7-8700 processor, 16GB memory, Nvidia GeForce GTX 1080Ti graphics card, and the deep learning framework uses PyTorch.

3.2. Evaluation Indicators
We choose commonly used some evaluation indicators to quantitatively evaluate the model, including precision (P), recall (R), mean average precision (mAP), and frames per second (FPS). Here

\[ P = \frac{TP}{FP+TP}, \quad R = \frac{TP}{TP+FN}. \]  

(3)

mAP is an index widely used in target detection, mainly used to evaluate the accuracy of the prediction frame. FPS can indicate the real-time performance of the target detection model. The higher the FPS value, the more pictures are processed per second, the faster the model inference speed, and the better the real-time performance.

3.3. Bolt Defect Detection Results
Compared with commonly used target detection algorithms, including Faster R-CNN, SSD, YOLOv3, and most recently ViT-FRCNN. The CNN backbone network of CViT-FRCNN is set to ResNet-18, and the final average pool used for classification is removed. Table 1 summarizes the final experimental data, and some test results are shown in figure 4. CViT-FRCNN obtained the highest precision, recall and mAP, and compared with the original ViT-FRCNN, it also has a larger performance improvement, which shows that extracting features via CNN will have more advantages than direct block embedding. In addition, CViT-FRCNN is also higher than other target detection models, thanks to the global attention of the self-attention mechanism.

Table 1. Experimental results.

| Method       | Precision | Recall | mAP  | FPS |
|--------------|-----------|--------|------|-----|
| Faster R-CNN | 85.5      | 86.2   | 82.4 | 10.3|
| SSD          | 85.1      | 87.1   | 83.7 | 17.2|
| YOLOv3       | 86.5      | 88.3   | 84.1 | 16.4|
| ViT-FRCNN    | 85.9      | 87.6   | 83.8 | 12.5|
| CViT-FRCNN   | **87.8**  | **89.9** | **85.2** | **11.5** |

Figure 4. Part of the test results.

However, experiments have also found that the FPS of the Transformer-based target detection model is lower than that of SSD and YOLOv3, and slightly higher than Faster R-CNN. This is because the computational complexity of the self-attention mechanism is the square of the input dimension [10]. Therefore, this limits the real-time performance of CViT-FRCNN to some extent. However, in the detection of the angle steel tower bolt defects, more attention is paid. Missing inspection may cause safety hazards and property losses, and even life-threatened. Therefore, in the case of meeting actual needs, it is a reasonable choice to sacrifice some real-time performance.
3.4. CNN Backbone Network
The influence of different CNN backbone networks on bolt defect detection was further tested, and the experimental results are shown in table 2. It is obvious that the residual series model is better than the VGG-16 directly stacked convolution, because the residual connection can enhance the information flow and reduce gradient disappearance [14]. In the residual model, ResNet-18 and DenseNet-121 have similar performance, but the model parameters, computational complexity and FPS of ResNet-18 are better than DenseNet-121, so ResNet-18 is more in line with the application scenarios of real-life angle steel tower inspection, thus achieving better performance overall.

Table 2. Experimental results.

| CNN backbone network | Precision | Recall | mAP  | FPS  |
|----------------------|-----------|--------|------|------|
| VGG-16               | 82.7      | 83.5   | 79.6 | 9.2  |
| DenseNet-121         | 88.1      | 89.6   | 84.8 | 8.5  |
| ResNeXt-29           | 87.5      | 88.3   | 84.2 | 9.8  |
| ResNet-18            | 88.3      | 89.4   | 85.2 | 11.5 |

3.5. Number of Transformer Encoders
This part studies the influence of the number of Transformer encoders on the bolt detection procedure. The experimental results are shown in table 3. The number of Transformer encoders reflects the ability of feature extraction, but experiments have found that rather than they are positively correlated, there exists a threshold. Which means that with the increase in the number of encoders, the performance is increased, while the FPS continues to decrease. However, performance saturation occurs after the number reaches 8, although there is still a slight performance improvement when the number reaches 10. Therefore, the number of encoders is 8 is a reasonable choice. When the number of encoders continues to increase, the performance begins to deteriorate, and over-fitting occurs, resulting in a decrease in the generalization performance of the model. Therefore, it is very important to set the number of Transformer encoders reasonably in practical applications.

Table 3. Experimental results.

| number of encoders | Precision | Recall | mAP  | FPS  |
|--------------------|-----------|--------|------|------|
| 4                  | 72.6      | 79.8   | 68.5 | 14.3 |
| 6                  | 85.6      | 84.4   | 80.4 | 12.6 |
| 8                  | 88.3      | 89.4   | 85.2 | 11.5 |
| 10                 | 88.7      | 89.1   | 85.7 | 10.1 |
| 12                 | 87.9      | 88.8   | 84.2 | 8.8  |

3.6. Residual Structure
Finally, table 4 shows an ablation experiment performed on the residual structure of CViT-FRCNN to verify the effectiveness of the module. After removing the residual structure, although the FPS has been slightly improved which saves the reasoning time of the residual block, other relevant comparison indicators have significantly decreased. Therefore, the residual module can further improve the detection performance of CViT-FRCNN, and it can ameliorate the problem of weakening of the information flow caused by the disappearance of the gradient. In addition, the size of the convolution kernel is set to 3×3, since the bolt target is small, the moderate scale of convolution kernel will not cause the loss of detailed information.

Table 4. Experimental results.

| Residual structure | Precision | Recall | mAP  | FPS  |
|--------------------|-----------|--------|------|------|
| ×                  | 86.7      | 88.1   | 83.9 | 11.8 |
| √                  | 87.8      | 89.9   | 85.2 | 11.5 |
4. Conclusions
In the natural environment, the angle steel tower bolts will appear rusty, loose and fall off, which makes the angle steel tower's stress situation change, and lays the bane of hidden dangers. Target detection algorithms based on deep learning can improve efficiency and reduce labor costs. This paper proposes a Transformer-based target detection network that uses a CNN backbone network to replace the direct block slicing operation of ViT-FRCNN to improve the richness of input features, and named it CViT-FRCNN. After that, a large number of experiments were carried out on the angle steel tower data set, and the best performance was achieved compared with the commonly used algorithms. In addition, the impact of different CNN backbones on the detection results was also tested. Results showed that the overall performance of the ResNet-18 backbone network is the best. Finally, the influence of the number of Transformer encoders and the residual structure on the detection performance is explored. Therefore, a large number of experimental results prove that the proposed CViT-FRCNN can meet the needs of actual detection.

Acknowledgement
This work was supported by the State Grid Anhui Electric Power Co., Ltd. (No. 52120019007G).

References
[1] Feng J J 2020 Implementation and research of automatic line inspection of power transmission lines based on 5G UAV Electric Power Equipment Management 9: 203-204
[2] Chen L, Chen D J 2021 Design of intelligent identification robot for insulator crack of power transmission line, Automation & Instrumentation 5: 197-201
[3] Zhao Z, Xu G, Qi Y, Liu N, Zhang T 2016 Multi-patch deep features for power line insulator status classification from aerial images International Joint Conference on Neural Network 3187-3194
[4] Tao X, Zhang D, Wang Z, Liu X, Zhang H, Xu D 2020 Detection of Power Line Insulator Defects Using Aerial Images Analyzed with Convolutional Neural Networks IEEE transactions on systems man and cybernetics 50(4): 1486-1498
[5] Girshick R, Donahue J, Darrell T, Malik J 2016 Region-Based Convolutional Networks for Accurate Object Detection and Segmentation IEEE transactions on pattern analysis and machine intelligence 38(1): 142-158
[6] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S E, Fu C Y, Berg A C 2016 SSD: Single Shot MultiBox Detector European Conference on Computer Vision 21-37
[7] Redmon J, Divvala S, Girshick R, Farhadi A 2016 You Only Look Once: Unified, Real-Time Object Detection Computer Vision and Pattern Recognition 779-788
[8] Dosovitskiy A, Beyer L, Kolesnikov A, et al. 2021 An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale International Conference on Learning Representations
[9] Vaswani A, Shazeer N, Parmar N, et al. 2017 Attention is All you Need Neural Information Processing Systems 5998-6008
[10] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers 213-229.
[11] Zhu X, Su W, Lu L, et al. 2020 Deformable detr: Deformable transformers for end-to-end object detection, arXiv preprint arXiv:2010.04159
[12] Beal J, Kim E, Tzeng E, et al. 2020 Toward transformer-based object detection arXiv preprint arXiv:2012.09958
[13] Wang F, Jiang M, Qian C, et al. 2017 Residual Attention Network for Image Classification Computer Vision and Pattern Recognition 6450-6458
[14] Bai J L, Kiros J R, Hinton G E 2016 Layer Normalization arXiv preprint arXiv:1607.06450
[15] Hendrycks D, Gimpel K 2016 Gaussian error linear units (gelus) arXiv preprint arXiv:1606.08415.