Fuzzy Systems and Data Mining IX
A.J. Tallón-Ballesteros and R. Beltrán-Barba (Eds.)

doi:10.3233/FAIA231076

DC-YOLOv5: Improved YOLOv5 for Transmission Line Fittings Detection Based on Deformable Convolution and Coordinate Attention

Qiang HE[a,1], Biaojun LI[a], Ting LIANG[a], Pei ZHOU[a], Mingquan YANG[a], Jun ZHU[a] and Ruiheng ZHOU[b]
[a] *China Southern Power Grid Co., Ltd. Extra High Voltage Transmission Company Guiyang Bureau*
[b] *North China Electric Power University*

Abstract. Transmission lines are an important component of the power system, and the detection of transmission line fittings is of great significance for ensuring the safe and stable operation of the power grid. In the inspection of transmission lines, drones are mainly used for taking photos and deep learning technology is used to achieve automatic detection. Due to the complex inspection background, high occlusion interference, and the variety of metal object categories and varying shapes and sizes, common detection methods have poor performance. This paper proposes an improved YOLOv5 method based on deformable convolution and coordinate attention, called DC-YOLOv5. Firstly, we construct the YOLOv5 network as the basic framework for the detection model. In order to extract more effective features from images containing complex background interference, we use deformable convolution to improve the original convolution module and enhance the feature extraction ability of the backbone network. Then, we use the coordinate attention module to process the output of the backbone network, improve the model's attention to fitting targets. This article hopes to effectively improve the performance of the model and maintain low complexity of the model for subsequent UAV deployment by using such uncomplicated lightweight modifications. Finally, in order to verify the effectiveness of DC-YOLOv5, a fitting detection dataset was established and experiments were conducted. The results indicate that DC-YOLOv5 has higher detection accuracy compared to other models and can accurately detect various metal object targets in complex environments.

Keywords. fittings detection; deformable convolution; attention mechanism; YOLOv5

1. Introduction

Transmission lines are an important component of emerging energy systems such as the energy internet and smart grids. Ensuring the stable operation of key components in transmission lines is the key to maintaining the stability of the power system, and is also an important part of the construction of the energy internet and smart grids[1]. Key

[1] Corresponding Author, Qiang HE, China Southern Power Grid Co., Ltd. Extra High Voltage Transmission Company Guiyang Bureau, Guiyang, China; Email: 2115049288@qq.com.

components such as fittings are widely used iron or aluminum metal accessories on transmission lines, mainly used for supporting, fixing, connecting bare wires, conductors, etc., including suspension clamps, grading rings, shockproof hammers, weights, etc. There are many types of fittings, and the shapes of different fittings also have certain differences. Due to being outdoors all year round and having complex contact with the environment, such components are likely to experience displacement, tilting, damage, etc. If these defects are not detected in a timely manner, they may lead to widespread power outages[2]. Therefore, automatic detection of fitting and prediction of faults are of great significance for ensuring the safe operation of the power grid[3].

With the maturity of drone technology, drone aerial photography technology has gradually replaced manual inspection. In recent years, with the advancement of deep learning, the use of computer vision technology to process aerial images of unmanned aerial vehicles on transmission lines and construct automated intelligent detection systems has become a current research hotspot. Reference [4] improved Faster R-CNN [5] as a component identification model for transmission lines, adjusted the size of convolutional kernels in convolutional operations, and expanded the dataset through data augmentation, verifying the feasibility of these two methods in improving accuracy. Reference [6] proposed an improved IoU (intersection over union) SSD [7] model for dense detection, which is more sensitive to target scale and adds repulsive loss to dense targets, achieving better dense detection results. There are many similar applications of computer vision technology in the field of transmission line detection.

The above research has to some extent achieved the detection of transmission line fittings, but there are still some problems. The inspection environment for power transmission lines is complex, with many interferences, and the shapes of different types of fitting vary greatly. The angles taken by drones are different, and conventional convolutional networks have weak adaptability to this, making it difficult to effectively extract available features. In order to solve these problems and improve the accuracy of transmission line fitting detection, this paper proposes an improved YOLOv5 transmission line fitting detection method based on deformable convolution [8] and coordinated attention [9] to address the problems in transmission line fitting detection. Our method is named DC-YOLOv5.

2. YOLOv5

2.1. Framework of YOLOv5

YOLOv5, like other algorithms in the series, is a typical one-stage object detection algorithm with a more efficient structure than two-stage algorithms such as Faster R-CNN. The structure of YOLOv5 framework is mainly composed of three parts: ① Backbone; ② Neck; ③ Detection Head The role of the backbone network is to extract features from input images and obtain feature layers for subsequent processing. The neck network is responsible for sampling and fusing feature layers at different scales, thereby enhancing the model's perception of targets at different scales, and effectively combining the shape information of large-scale feature layers with the semantic information of small-scale feature layers. The detection head will process the processed feature layers and predict the type and position information of different targets in the image.

2.2. Specific composition

In YOLOv5, the backbone adopts an improved CSPDarknet network. In the backbone network, feature extraction mainly relies on convolution, and the convolutional blocks in CSP Marknet are composed of convolution, batch normalization, and activation functions. In order to increase the depth of the network while avoiding the problem of gradient vanishing during the training process, CSPDarknet adopted the residual idea of ResNet [10] and constructed a C3 module, which utilizes channel dimensionality reduction and dimensionality increase to increase the receptive field of the model, facilitating the extraction of more detailed features.

The YOLOv5 neck network adopts a PAFPN [11] structure. Compared to traditional feature pyramid networks, PAFPN adds a reverse downsampling path, combining semantic features transmitted from top to bottom and shape features transmitted from bottom to top, enhancing the aggregation ability of the network and ensuring accurate prediction of images of different sizes. Similarly, YOLOv5 has added a C3 module to the neck network PAFPN, allowing the model to learn more features. Finally, the feature layer will be fed into the detection head, predicted by the model, and the loss will be calculated for training. The loss function of YOLOv5 mainly consists of three parts, including basic classification loss and regression loss, as well as confidence loss. The three parts of the loss will be summed after weighting to obtain the total loss, which will be used for backpropagation and model training.

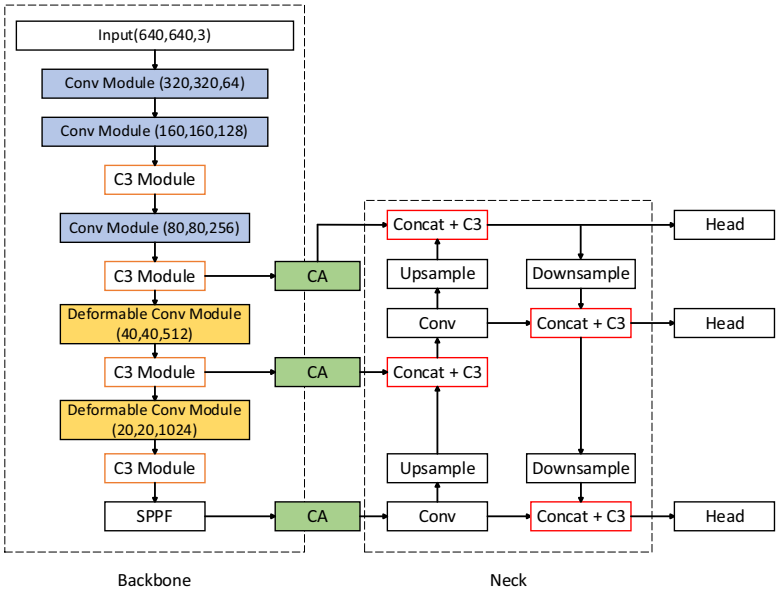


Figure 1. The structure of DC-YOLOv5

3. DC-YOLOv5

The YOLOv5 model can achieve good results in object detection and relies on its lightweight characteristics, making it very friendly for industrial deployment. This paper

proposes an improved YOLOv5 transmission line fitting detection model called DC-YOLOv5. Its structure is shown in Figure 1. DC-YOLOv5 has been mainly improved and expanded in two aspects: (1) using deformable convolution to improve partial convolution operations in the original backbone network. Deformable convolution enhances the model's feature extraction ability for targets of different scales by adding irregular offsets to the convolution. (2) Add coordinate attention mechanism to the network to make the model pay more attention to useful features of images outside the background and improve detection accuracy.

3.1. Deformable convolution

In order to enhance the feature extraction ability of the backbone network for targets of different scales and irregular targets, we use deformable convolution to replace some ordinary convolutions in the YOLOv5 backbone network. Figure 2 shows the difference between standard convolution kernels and deformable convolution kernels.

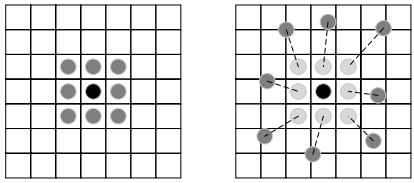


Figure 2. The structure of deformable convolution

The left figure shows the standard convolutional kernel, which slides at a fixed size during operation; The figure on the right shows the deformable convolution kernel with learnable offsets added. Deformable convolution kernel is not constrained by a fixed size near the sampling point position and is trained to learn how to set the optimal offset without additional supervision. Deformable convolution enhances the network's feature extraction ability for targets of different scales through sampling at irregular positions.

3.2. Coordinate attention mechanism

In order to enhance the features extracted by the model in complex backgrounds and focus the model's attention on key features, we use coordinate attention mechanism to improve the backbone output and improve the accuracy of detection. The structure of the coordinate attention mechanism is shown in Figure 3.

The calculation of coordinate attention not only considers the relationship between channels, but also considers the position information in the direction of the feature space, and is lightweight enough to not increase too much computational overhead. Firstly, the input feature map will undergo global average pooling in both the width and height directions to obtain features in both directions, $Z^h \in \mathbf{R}^{\mathrm{C}\times\mathrm{H}\times 1}$ and $Z^w \in \mathbf{R}^{\mathrm{C}\times 1\times\mathrm{W}}$.Then, the two directional feature maps obtained from the global receptive field are concatenated and fed into the shared 1 × 1 convolution module, which reduces the dimensionality of the channel by *r* times and sends it into the normalization layer and activation function to obtain the feature layer f, $f \in \mathbf{R}^{\mathrm{C/r}\times 1\times(\mathrm{H+W})}$. Then decompose f again along the dimensions of height and width to obtain $f^h \in \mathbf{R}^{\mathrm{C/r}\times\mathrm{H}\times 1}$ and $f^w \in \mathbf{R}^{\mathrm{C/r}\times 1\times\mathrm{W}}$, and separately utilize the other 1 × 1 convolution module to adjust the

convolution to the original number of channels and activate it using the sigmoid activation function to obtain the attention weights g^h in the height direction and g^w in the width direction of the feature map, respectively. After the above calculation, the attention weights of the input feature map in the high and wide directions will be obtained. Finally, by multiplying and weighting the height and width of the original feature map, the final feature map with attention weight will be obtained, which improves the model's attention to effective features.

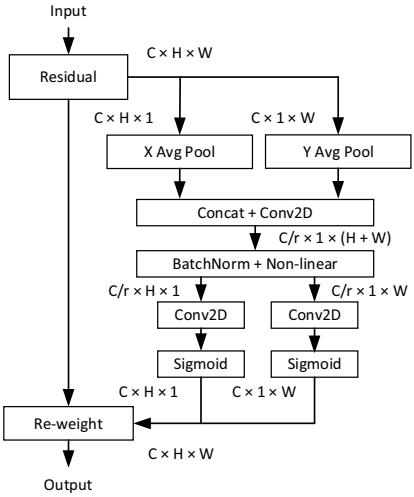


Figure 3. The structure of coordinate attention mechanism

4. Experiments

4.1. Dataset

This paper collects aerial images of unmanned aerial vehicles and constructs a transmission line fitting detection dataset after data screening. The dataset includes 12 types of fittings, including suspension clamps, grading rings, shielded rings, shockproof hammers, and includes 1586 images and 8329 annotation objects. We divided the dataset into a training set and a validation set in a 4:1 ratio for experimental purposes. Some example images of the dataset are shown in Figure 4. The detailed dataset composition is shown in Table 1.

(a)Example 1

(b)Example 2

Figure 4. Dataset examples

Table 1. Transmission Line Fitting Detection Dataset

Fittings	Nums	Fittings	Nums
pre-twisted suspension clamp	160	shielding ring	141
bag-type suspension clamp	986	grading ring	726
shockproof hammer	1213	u-type hanging ring	1325
yoke plate	832	wedge-tpye strain clamp	105
adjusting plate	619	weight	357
hanging board	1524	spacer	341

4.2. Experimental Results and Analysis

In the experiment, we compare our model with several object detection models on the dataset. To demonstrate the superiority of our method, we use the metric mAP in object detection to measure model performance. The AP^{50} indicator is relatively broad while the AP^{50-95} indicator is more stringent. The experimental results are shown in Table 2.

Table 2. Comparison of different models

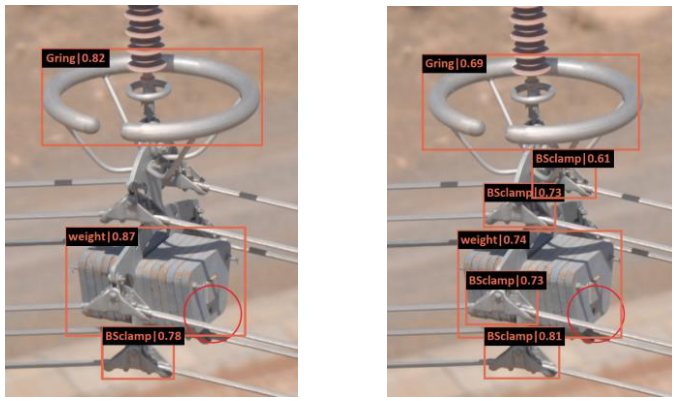
Method	mAP(%)		FPS
	AP^{50}	AP^{50-95}	
SSD[12]	70.54	49.27	43.7
Faster R-CNN	76.32	52.33	26.8
RetinaNet[13]	74.18	50.41	31.6
YOLOv5	76.25	52.36	82.6
DC-YOLOv5(Ours)	**80.31**	**54.52**	79.2

The experiment compared our proposed method with models such as SSD, RetinaNet, original YOLOv5, Faster R-CNN, etc. Compared with Faster R-CNN, the one-stage detection models SSD and RetinaNet have improved inference speed, but their accuracy is slightly lower. Our DC-YOLOv5 model is a one-stage detection model, but it has improved accuracy and detection speed compared to them, and its indicators exceed the two-stage model Faster R-CNN, demonstrating good detection performance. Overall, our method has a significantly faster inference speed and higher effectiveness than traditional models. Compared with the original YOLOv5 model, DC-YOLOv5 has an increase of 4.06% in AP^{50} and 2.16% in AP^{50-95}, resulting in a certain improvement in detection accuracy. Although deformable convolution and coordinated attention mechanism sacrifice some computational effectiveness and slightly reduces FPS, it is still acceptable. Overall, DC-YOLOv5 performs the best in the detection.

To further validate the effectiveness of the proposed improvement method on the model, we conducted ablation experiments on the model, and the experimental results are shown in Table 3. After introducing deformable convolution into the backbone of the baseline model, the accuracy improved by 1.91%; After using the coordinate attention mechanism in the baseline model, the accuracy improved by 2.38%. After combining these two improved methods, DC-YOLOv5 achieved the optimal detection accuracy, increasing AP^{50} to 80.31%, an increase of 4.06%, proving the effectiveness of the proposed method. Due to the introduction of additional parameters and modules in both of our improved methods, the FPS of the model decreased slightly but it is acceptable.

Table 3. Ablation study

Method		AP^{50}(%)	FPS
Deformable convolution	Coordinate attention		
		76.25	82.6
√		78.16	80.2
	√	78.63	81.4
√	√	**80.31**	**79.2**



(a)YOLOv5 (b)DC-YOLOv5

Figure 5. Comparison of detection results between YOLOv5 and DC-YOLOv5

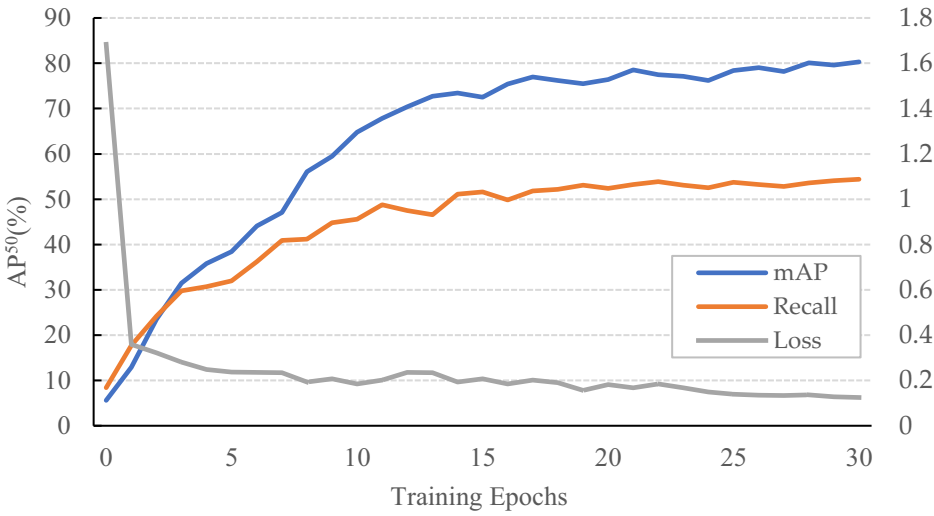


Figure 6. Different metrics in the training process

Finally, we conducted visualization experiments. The results are shown in Figure 5. Figure 5 (a) shows the result of original YOLOv5, and Figure 5 (b) shows the result of DC-YOLOv5. From the detection results, it can be seen that in the middle of the image, due to the complex occlusion relationship, the original YOLOv5 missed three suspension clamps, resulting in missed detection. The improved DC-YOLOv5 successfully detected these three easily missed clamps, and the detection effect was better than the baseline YOLOv5, proving the effectiveness of the improved method in this paper. In addition, Figure 6 shows the metrics of our model in training epochs such as AP^{50}, Recall and Loss, demonstrating the stability of the improved model during the training process.

5. Conclusion

Using computer vision technology to process UAV aerial images on transmission lines and constructing automated intelligent detection systems has become a research hotspot

in recent years. This paper proposes DC-YOLOv5, which uses deformable convolution and coordinate attention mechanism to extract more effective features in images with complex background interference and allow the model to focus on key features and improve detection accuracy. Experiments on the transmission line fitting detection dataset show that DC-YOLOv5 can accurately detect multiple fitting targets in complex environments, and has higher detection accuracy compared to other models.

In further work, we plan to deploy our lightweight model on UAVs with edge intelligent devices and achieve real-time detection during UAV inspections. We hope our method can make a contribution to ensuring the stable operation of the power grid.

References

[1] DONG Zhaoyang，ZHAO Junhua，WEN Fushuan，et al. From smart grid to energy internet：basic concept and research framework[J]. Automation of Electric Power Systems，2014，38(15)：1-11.

[2] Li L. The UAV intelligent inspection of transmission lines. In Proceedings of the 2015 International Conference on Advances in Mechanical Engineering and Industrial Informatics. Atlantis Press, 2015; pp.1542-1545.

[3] ZHAO Zhenbing，JIANG Zhigang，LI Yanxu，et al. Overview of visual defect detection of transmission line components[J]. Journal of Image and Graphics，2021，26(11)：2545-2560.

[4] Tang Yong, Han Jun, Wei Wenli, et al Research on part recognition and defect detection of transmission line in deep learning[J]. Electronic Measurement Technology, 2018, 41(6): 60-65.

[5] Ren S, He K, Girshick R, et al Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28.

[6] Qi Yincheng, Jiang Aixue, Zhao Zhenbing, et al Detection method of transmission line inspection image fittings based on improved SSD model [J]. Electrical Measurement & Instrumentation, 2019, 56(22): 7-12.

[7] Liu W, Anguelov D, Erhan D, et al Ssd: Single shot multibox detector[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016: 21-37.

[8] Dai J, Qi H, Xiong Y, et al. Deformable convolutional networks[C]//Proceedings of the IEEE international conference on computer vision. 2017: 764-773.

[9] Hou Q, Zhou D, Feng J. Coordinate attention for efficient mobile network design[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 13713-13722.

[10] He K, Zhang X, Ren S, et al Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[11] Liu S, Qi L, Qin H, et al. Path aggregation network for instance segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 8759-8768

[12] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016: 21-37.

[13] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.