Power Grid Field Violation Recognition Algorithm Based on Enhanced YOLOv5

Wenwen Zhu\textsuperscript{1a}, Yuanzhong Shu\textsuperscript{1b*}, Shuo Liu\textsuperscript{2a}

\textsuperscript{1}College of Information Engineering, Nanchang Hangkong University, Nanchang, Jiangxi, 330063, China
\textsuperscript{2}College of Electrical Engineering, Anhui Polytechnic University, Wuhu, Anhui, 241000, China

\textsuperscript{1b*}36012@nchu.edu.cn, \textsuperscript{1a}1904081200104@stu.nchu.edu.cn, \textsuperscript{2a}lindsayshuo@foxmail.com

Abstract—Aiming at the problem that Yolov5 is difficult to accurately detect backlighting and small samples in national grid power transmission and distribution operations, studies the illegal behavior technology of “two wear and one wear” in power grid operation, and proposes a violation recognition algorithm based on the enhanced YOLOv5. First, add a new detection layer and use the BiFPN (bi-directional feature pyramid network) layer for feature fusion, so that feature layers of different scales can better learn the weight distribution and enhance the fusion ability. Secondly, add CBAM (Convolutional Block Attention Module) module before the output detection layer feature map to make full use of channel and spatial information to achieve better model accuracy and recall. The experimental results on the six self-made data sets show that the mAP of this method is 91.0%, which is 5.6% higher than the original algorithm, indicating that the model has stronger predictive ability and robustness for the identification of personnel and safety tools in transmission and distribution scenarios.

1. Introduction

The power grid operation site has large space, high risk, and difficult management and control. In recent years, the convolutional neural network model has developed rapidly and has been widely applied\cite{1,2}. Many scholars have published a large number of excellent models based on deep learning\cite{3}. It focuses on the intelligent control of on-site operation safety, and timely reminds the operation risks and automatically alarms based on the recognition of violations in the grid operation site. It is divided into Two-Stage algorithm and one-stage detection algorithm mainly according to whether there is candidate box generation. The former has high detection accuracy, large computation and slow speed, which is difficult to meet the needs of real-time target detection. The latter has the advantages of less computation, faster speed, and stronger real-time performance and practicability. For example, the R-CNN series methods proposed by Girshick\cite{4,5,6} et al., such two-stage method firstly generates candidate regions based on feature extraction, and then determines the target category and position by regression of candidate regions. The detection rate can reach 5fps.

The SSD (Single Shot multibox Detector) method proposed by Liu et al. \cite{7}, the target detection algorithm based on YOLO (You Only look once) proposed by Redmon et al. \cite{8,9,10}, V2 uses darknet-19 network model to improve mAP value from 63.4% to 78.6%, while V3 uses the new Darknet-53 as feature extractor, which has the same accuracy as SSD while 3 times faster, and also
significantly improves detection of small objects. Alexey et al. [11] proposed the YOLOv4 algorithm, in which CSPDarknet53 draws on the network structure of cross-stage partial network (CSPnet) and SPP, as well as data enhancement using the Mosaic method. The aforementioned One-Stage algorithm directly predicts the coordinates and category of the bounding box based on the features extracted by the network, and the detection speed reaches 45fps. Literature[12] reduces the dimensionality of the convolutional layer of the YOLOv5 network and improves the post-processing method, which improves the detection performance of the quality of inkjet characters. Literature[13] improves YOLOv5 as a lightweight backbone network to meet faster and more accurate vehicles detection.

This paper proposes a recognition algorithm of illegal behavior based on improved YOLOv5, aiming at the problems of difficult supervision of power grid construction site, high personnel cost and difficult judgment of illegal behavior. The specific improvement scheme is as follows: add multi-scale detection layer for small targets in YOLOv5 network under pytorch framework and use BiFPN[14] connection to make four detection layers learn important features and use CBAM[15] attention. The performance of small target detection is improved, and the sensitivity of the model to background noise is reduced, so that the model for identifying violation behaviors achieves a higher degree of fit. Finally, the practicability and robustness of the proposed method are verified by experiments.

2. Violation Recognition Based on Improved YOLOv5 Algorithm

2.1. YOLOv5 algorithm principle

YOLOv5 is similar to the channel and layer control factors of EfficientNet [16], which can flexibly configure models of different complexity. According to the model depth multiple and layer channel multiple, it can be divided into four types: YOLOv5s, YOLOv5m, YOLOv5l, YOLOv5x. The YOLOv5 network structure is shown in Fig.1.

![YOLOv5 network overall structure diagram](image)

Fig.1 YOLOv5 network overall structure diagram

This article uses the smallest model YOLOv5s, the depth multiple is 0.33, the width multiple is 0.50, the number of BottleneckCSP (True) is 1, 3, 3, the number of BottleneckCSP (False) is 1, and the number of convolution kernels is 32, 64, 128, 256, 512, the fastest and easy to deploy quickly.

Input uses Mosaic data enhancement, randomly using 4 pictures, random zooming, random cropping and random arrangement for splicing, enriching the data set, and significantly improving the
recognition ability of small objects. The data of 4 pictures can be calculated at the same time, without a large Mini-batch, which reduces the consumption of memory. The use of k-means clustering to calculate the optimal anchor value of the data set and the use of adaptive image scaling reduce the black edges and improve the detection speed. Backbone mainly adopts FOCUS, CSPnet and SPP structure, as shown in Fig.2. Focus performs 4 slices, and then passes through a convolutional layer with a convolution kernel number of 32, turning the initial 3x640x640 image into a 32x320x320 feature map. It is used to reduce the amount of calculation and the number of layers to increase the speed, and at the same time minimize the loss of information for down-sampling. CSPnet first divides the feature mapping of the base layer into two parts, and then splices the two parts through a cross-stage hierarchical structure, which reduces the amount of calculation while ensuring accuracy. SPP uses 4 convolution kernels of different sizes for maximum pooling, and the number of hierarchically fixed windows, each window corresponds to a layer of the pyramid, and the window size is adjusted to achieve multi-level pooling and fixed output.

![Fig.2 Backbone component structure diagram](image)

Head uses feature pyramid networks (FPN) and Path Aggregation Network (PAN). FPN transfers features from top to bottom to solve multi-scale problems. PAN aggregates shallow features from the bottom-up path, and different feature layers are fully integrated, which speeds up the flow of information. Output uses the binary cross-entropy loss function to calculate the category probability and target confidence score loss. When calculating the bounding box loss, GIOU_Loss [17] is used, as shown in Equation 1:

$$ L_{GIOU} = 1 - IoU + \frac{A_c - U}{A_c} $$

Among them, $IoU$ (Intersection over Union) is the intersection ratio of the real box and the predicted box, $A_c$ is the minimum closure area of the real box and the predicted box, and $U$ is the intersection of the real box and the predicted box.

Through real-time streaming protocol (RTSP) access to outdoor job site cameras and indoor webcam video stream, combined with YOLOv5 algorithm for real-time analysis, to identify wearing helmets, wearing work clothes and safety belts. YOLOv5 has better detection results and real-time
performance. However, in the outdoor high-altitude operation scene in this article, the helmet and safety belt targets under the long-range lens are often small. The original down-sampling 80*80, 40*40, 20*20 three This feature layer is difficult to detect targets smaller than 8*8, that is, there will be a certain degree of missed detection and false detection. In the original Head branch, features of different scales are equally fused, and the importance of feature layers of different scales is not considered.

2.2. Improved multi-scale feature layer

It can be seen from Fig.1 that YOLOv5 performs 3, 4, and 5 down-sampling of pictures with an input of 640*640 to obtain three detection layers of 80*80, 40*40, and 20*20, which are used to detect small, medium and large targets of different scales. Considering that the detected helmets and safety belt cameras are mostly small in the long view, which is easy to cause missed detection and misdetection, a feature layer for detecting smaller targets is added. As shown in Fig.3, the first step is to add BottleneckCSP and CONV to extract features on the basis of the original 16 layers, and then use Upsample to up-sampling to obtain a feature map of 160*160. In the 20th layer and the second layer of the backbone network, 160* 160 feature maps are fused. In the second step, we continue to use BottleneckCSP, firstly obtain the feature map of the 22nd layer for predicting the smaller target through convolution, and then down-sampling by the CONV module with a step size of 2 to obtain the 80*80 feature map. In the 24th layer and The 80*80 feature map of the 18th layer is fused. After completing the improvement of the multi-scale feature layer through the above two steps, it can be found that the model now uses four-scale feature maps for prediction, making full use of the resolution of the bottom layer and the semantic features of the high layer.

![Fig.3 Added multi-scale feature layer structure map](image)

2.3. Add feature fusion of BiFPN and CBAM

In this paper, the BiFPN layer is used to replace the original tensor splicing Concat layer, which corresponds to the 12th, 16th, and 20th layers in the upsampling and the 24th, 28th, and 32th layers in the downsampling, all replaced with BiFPN. Combining the down-sampling features of the backbone network with the up-sampling features in the Head, and then fusing the down-sampling features in the Head, so that the network has stronger semantic information, thereby solving the missed detection caused by small targets and repeated feature extraction in backlight scenes As well as the problem of misdetection, the model has a better feature extraction effect on targets with fuzzy features. The improved feature fusion and BiFPN structure are shown in Fig.4.
Mainly adopt two-way cross-scale connection and fast normalization fusion. BiFPN removes nodes with only one input edge. This node has very little contribution to the feature network that integrates different features. Add additional edges at the same level of input to output nodes to fuse more features without adding too much cost. And repeat each bidirectional path as a feature network layer multiple times to achieve a higher level of feature fusion. The formulas for fast normalized fusion and BiFPN are shown in equations 2 and 3.

\[
O = \sum_i \frac{w_i}{\varepsilon + \sum_j w_j} \cdot I_i, w_i \geq 0
\]

\[
P_{6}^{\text{td}} = \text{Conv}(\frac{\frac{w_1}{\varepsilon} \cdot P_{6}^{\text{in}} + \frac{w_2}{\varepsilon} \cdot \text{Resize}(P_{7}^{\text{in}})}{1 + \frac{w_1 + w_2}{\varepsilon}})
\]

\[
P_{6}^{\text{out}} = \text{Conv}(\frac{\frac{w_1}{\varepsilon} \cdot P_{6}^{\text{in}} + \frac{w_2}{\varepsilon} \cdot P_{6}^{\text{td}} + \frac{w_3}{\varepsilon} \cdot \text{Resize}(P_{5}^{\text{out}})}{1 + \frac{w_1 + w_2 + w_3}{\varepsilon}})
\]

In formula (2), \(w_i\) is a learnable weight, and the Relu activation function is used after each \(w_i\). The normalized weight is between 0 and 1, and the \(\varepsilon\) value is 0.0001, which is used to avoid numerical instability. Equation (3) takes the 6th level as an example, \(P_{6}^{\text{td}}\) is the intermediate feature of the 6th level on the top-down path, \(P_{6}^{\text{out}}\) is the output feature of the 5th and 6th levels in the bottom-up path, \(P_{6}^{\text{in}}, P_{7}^{\text{in}}\) is the input of the 6th and 7th levels feature. Resize is usually up-sampling or down-sampling, and Conv is usually a convolution operation of feature processing.

The convolutional attention module CBAM makes the model focus on key areas through channels and spaces, enhances important features in the feature map, and suppresses useless features. It can be divided into a channel attention module and a spatial attention module. The channel attention module
is shown in Fig.5(a), which performs maximum pooling and average pooling on the input feature map. Through sharing MLP, the activation function is used to output the channel attention after addition, and then multiplying to get the final channel attention. The spatial attention module is shown in Fig.5(b). After the feature map of the channel attention module is obtained, the maximum pooling and average pooling are performed respectively, and the spatial attention is obtained through convolution and activation after splicing, and then the channel attention The force feature map is multiplied to obtain the final output feature map.

Consider the following ways to add CBAM modules: The first is to add the last layer at the end of the backbone network extraction of features, the second is to add after the BottleneckCSP in the head sub-sampling stage, and the third is to integrate the first two methods. The experiment adopts the second scheme, which is mainly to add CBAM fusion before convolution and output 4 detection feature layers in the Head branch, namely the 22nd, 27th, 32nd, and 37th layers.

Through the above-mentioned feature fusion operation, the expression ability of the target feature is better, the target has better robustness, and the recognition accuracy of the network is improved. After labeling training on 12088 samples and testing on 2418 test sets, it is found that the recall rate, accuracy rate and mAP of the improved network are significantly increased.

3. Experiment

3.1. Data set and experimental method

According to Pascal VOC[18]2007 data set format, video data were obtained from cameras in various complex environments such as power transmission and distribution lines and substations, and labeling tool was used to mark categories and locations. Compliance behavior refers to the correct wearing of safety helmet and standard wearing of overalls and safety belts with the corresponding labels of helmet, workcloth and belt. Violation behavior mainly refers to the improper wearing of safety helmet, overalls and safety belts with the corresponding labels of head, person and nobelt. A data set containing 12088 pictures has been established. And use a script to convert it into a txt file required to train the yolo model.
Add a multi-scale detection layer to the YOLOv5s network, use BiFPN and CBAM modules for feature fusion, improve the traditional YOLOv5s, and use the same data set for ablation comparison experiments. The experiment uses GeForce RTX3080 graphics card, Ubuntu 18.04 host with 10G video memory, based on pytorch 1.7.1 framework, configuration cuda11.0, using python language to improve the training of the YOLOv5s network model by migration learning on the self-made data set, and set the batch size to 8. The learning rate is 0.01 and the epoch is 300.

In order to measure the real-time performance of the algorithm, FPS (Frame Per Second) is used to test the detection speed, that is, the number of pictures detected per second. In addition, the problem studied in this paper belongs to multiple classifications, so mAP is introduced to evaluate the improved model. mAP is the sum of the average accuracy of all categories divided by all categories. The higher the mAP, the better the classifier. As a relatively good evaluation index, mAP is calculated as shown in formulas (4) - (7):

$$AP = \frac{1}{101} \sum_{r \in \{0,0.01,...,1\}} p_{\text{interp}}^{r}(r)$$  \hspace{1cm} (4)

$$p_{\text{interp}}^{r}(r) = \max_{\tilde{r} \in r} p(\tilde{r})$$ \hspace{1cm} (5)

$$p = \frac{TP}{TP + FP}$$ \hspace{1cm} (6)

$$r = \frac{TP}{TP + FN}$$ \hspace{1cm} (7)

Among them: TP is a positive sample for correct prediction, TN is a negative sample for correct prediction, FP is a negative sample for incorrect prediction, and FN is a positive sample for incorrect prediction. $p$ (precision) is the accuracy rate, $r$ (recall) is the recall rate. After the training, compare the recognition effect of the model before and after the improvement, and measure the performance index of the algorithm.

3.2. Experimental results

During the experiment, the changes of location loss, classification loss and confidence loss are shown in Fig.6. The abscissa is the training batch, the ordinate is the loss value, GIoU Loss predicts the error between the bounding box and the real box, the Objectness Loss predicts the target of the box, and the Classification Loss corresponds to the loss of the classification. It can be seen that the curve of the loss function drops rapidly in the first 50 epochs, and tends to balance after 250 epochs, and the model converges.

![Fig.6 Loss function graph](image-url)
The IoU threshold set in the experiment is 0.5. When IoU>0.5, the sample is considered to be a positive sample, otherwise the sample is considered to be a negative sample. The improved YOLOv5 algorithm performs two wears (tooling, safety belt) and one wears a (helmet) to detect compliance and violations, and the recognition results of the improved algorithm and the initial algorithm are shown in Fig.7.

Fig.7 Network recognition results before and after improvements

For dense scenes and pictures in the long view of the camera, YOLOv5 has missed and repeated detections for small targets wearing helmets. As shown on the left side of Fig.7, small targets wearing helmets and violations without helmets all have leaks. In addition, the occluded tooling is not recognized. After the addition of multi-scale feature layers, the detection of small targets is more accurate. The improved network of feature fusion detects both compliance and violations of the helmet wearing. Tooling is also accurately identified. In addition, in the backlit scene under high-altitude operations in the second column, the confidence in the detection of helmets and safety belts is also higher.

3.3. Comparative experiment analysis
Use 2418 test set images for testing. Method 1 adds a multi-scale feature layer to YOLOv5, Method 2 uses BiFPN layer for feature fusion based on Method 1, and Method 3 adds CBAM on the basis of 2, to get mAP and other performance before and after improvement. The indicators are shown in Table 1:

| Method     | Multi-scale feature layer | BiFPN | CBAM | mAP(%) | recall(%) | FPS   |
|------------|---------------------------|-------|------|--------|-----------|-------|
| YOLOv5     |                           |       |      | 85.4   | 93.0      | 90.08 |
| Method 1   | √                         |       |      | 86.1   | 93.6      | 80.80 |
| Method 2   | √                         | √     |      | 86.5   | 94.1      | 78.07 |
| Method 3   | √                         | √     | √    | 91.0   | 95.0      | 77.39 |
As can be seen from Table 1, the mAP of feature layer used for detecting small targets in Method 1 is slightly improved. Compared with the original YOLOv5, the mAP and recall of Method 3 are improved by 5.6% and 2.0% respectively, that is, the network is more excellent and the generalization ability is also improved, but the value of FPS is decreased to some extent. Considering that the algorithm has met the real-time demand, the calculation cost is sacrificed for higher accuracy. For Method 3, add attention module to Head branch, method 4 adds attention module to backbone network during feature extraction, and method 5 adds attention module to backbone network and Head branch at the same time. The test results are shown in Table 2.

| Method   | Backbone | Head | mAP(%) |
|----------|----------|------|--------|
| Method 3 | √        |      | 91.0   |
| Method 4 | √        |      | 90.6   |
| Method 5 | √        | √    | 89.8   |

The mAP value of the attention module used in Method 3 in Table 2 is slightly higher than that in Method 4, which is less than the attention module added in Method 5, and achieves a higher accuracy rate. In summary, finally adopt method 3 as the final improvement network, and train on the VOC07+12 data set to test the performance of the algorithm before and after the improvement. As shown in Table 3, the VOC data set is of higher quality than the data set obtained from the camera in this article, Method 3 increases the mAP value by 9.3% and 6% compared with the original method, and the network performance is better.

| Method   | mAP(%) | recall(%) |
|----------|--------|-----------|
| YOLOv5   | 68.8   | 90.0      |
| Method 3 | 78.1   | 96.0      |

4. Conclusion
It is difficult for the power grid safety supervisor to judge whether the construction personnel are wearing safety hats and safety belts and whether the number of construction personnel meets the demand through the camera at the operation site. In this paper, multi-scale feature layer is added to YOLOv5s network, and BiFPN and attention CBAM modules are improved for feature fusion, and a violation recognition algorithm based on enhanced YOLOv5 is proposed. The above problems are solved, and it is also convenient for the management of safety supervision personnel, and provides a guarantee for the identification of violations in power transmission and distribution scenarios.

Acknowledgments
This work was financially supported by the National Natural Science Foundation of China (71761028, 61572032); the Science and Technology Research Project of Jiangxi Provincial Department of Education (209924); the University Synergy Innovation Program of Anhui Province (GXXT-2020-070).
References

[1] YAN C M, WANG C. (2021) Development and application of convolutional neural network model. J. Journal of Frontiers of Computer Science and Technology, 15(01): 27-46.

[2] LIU S, GE Y, LI Y Y, et al. (2021) Recognition Algorithm of Unmanned Ship at Sea Based on Improved SSD. J. Transducer and Microsystem Technologies, 40(02): 120-123.

[3] GAO Y, XIAO D. (2020) Small object detection algorithm based on multi-feature fusion. J. Computer Engineering and Design, 41(07): 1905-1909.

[4] Girshick R, Donahue J, Darrell T, et al. (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Columbus. pp. 580–587.

[5] Girshick R. (2015) Fast r-cnn. In: IEEE International Conference on Computer Vision. Santiago. pp. 1440–1448.

[6] REN S Q, HE K M, Girshick R, et al. (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. J. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6): 1137–1149.

[7] LIU W, Anguelov D, Erhan D, et al. (2016) SSD: Single shot multibox detector. In: European Conference on Computer Vision, Amsterdam. pp. 21–37.

[8] Redmon J, Divvala S, Girshick R, et al. (2016) You only look once: Unified, real-time object detection. In: IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas. pp. 779–788.

[9] Redmon J, Farhadi A. (2017) YOLO9000: Better, faster, stronger. In: IEEE Conference on Computer Vision and Pattern Recognition. Honolulu. pp. 6517–6525.

[10] Redmon J, Farhadi A. (2018) YOLOv3: An incremental improvement. J. arXiv: 1804. 02767, 2018.

[11] Bochkovskiy A, Wang C Y, Liao H M. (2020) YOLOv4: Optimal speed and accuracy of object detection. J. arXiv preprint arXiv: 2004. 10934, 2020.

[12] GE Q, LIANG Q K, ZOU K L, et al. (2021) Detection Method of Inkjet Code Quality Based on Lightweight Network and Embedded. J/OL. Control Engineering of China, 1-9[2021-11-04]. https://doi.org/10.14107/j.cnki.kzgc.20200903.

[13] WANG C P, WANG H Q, YU F J, et al. (2021) A high-precision fast smoky vehicle detection method based on improved Yolov5 network. In: IEEE International Conference on Artificial Intelligence and Industrial Design (AIID). Guangzhou. pp. 255-259.

[14] TAN M X, PANG R M, LE Q. (2020) EfficientDet: scalable and efficient object detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle. pp. 10778-10787.

[15] Woo S, Park J, Lee J-Y, et al. (2018) Cham: convolutional block attention module. In: European conference on computer vision(ECCV). Munich. pp. 3-19

[16] TAN M, LE Q. (2019) Effiicientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. Long Beach. pp. 6105-6114.

[17] Rezatofighi H, Tsoi N, Gwak J, et al. (2019) Generalized intersection over union: a metric and a loss for bounding box regression. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach. pp. 658–666.

[18] Everingham M, Gool L, Williams C K, et al. (2010) The Pascal visual object classes (VOC) challenge. J. International Journal of Computer Vision, 88(2): 303–338