A smoking behavior detection method based on the YOLOv5 network

A smoking behavior detection method based on the YOLOv5 network

Xiangkui Jiang[1], Haochang Hu[1,*], Xun Liu[1], Rui Ding[1], Yuanbo Xu[1], Jianxu Shi[1], Yaoyao Du[1] and Chunlin Da[1]

[1]School of Automation, Xi'an University of Posts and Telecommunications
Xi'an, China

hu15194438027@163.com

Abstract. Smoking in public places not only brings about some safety hazards, but also does harm to people's lives, property and living environment. A smoking behavior detection model based on deep learning is trained for the concern of environment and safety. First, a vertical rotation data enhancement method is adopted in the preprocessing stage to extend the dataset and increase the objects of detection. Then, the channel attention module is introduced in backbone network to calibrate the feature response. Finally, added a small target detection layer to the YOLOv5 algorithm. This paper analyzes the network structure of the YOLOv5s, and the model is trained and tested by utilizing the YOLOv5s network. Experimental results show that the mAP value of the algorithm is improved by 5.3% over the original algorithm.

1. Introduction

Smoking has become a worldwide public health problem which is hard to solve. Harms caused by smoking is well acknowledged, and they can indirectly or directly lead to many diseases and even threaten people's lives. According to the World Health Organization, cigarette smoking may cause premature death. Smoking has attracted widespread concerns in the areas of fire prevention and public safety, environment, and health as well as other aspects in behavioral research[1-3].

According to the different detection methods, smoking detection can be classified into physical equipment detection methods and smoking detection technology which is based on computer vision. The traditional physical device detection method utilizes smoke sensors and wearable devices etc[4]. In traditional graphics, the object detection method is not ideal, since it is prone to be disturbed by other objects. Besides, the method also tends to result to inaccurate positioning. Secondly, the method requires a large amount of calculations, and needs to process the features of and to judge the classification of each sliding window. Finally, the approach and the process of manual feature-extraction are complicated without generalization.

Jianhao Liao et al.used smoking targets were detected based on YOLOv3 with a mAP value of 0.76[5]. Weibin Cai et al.used mosaic data augmentation to avoid model overfitting which is caused by the simple background in the public datasets and the YOLO-SMOKE model outperforms the original model by 4.91% mAP[6]. Rentao Zhao et al. proposed an improved algorithm based on YOLOv3-tiny deep learning network for indoor smoking behavior detection, which can effectively meet the practical application requirements and provide a new way for assisting indoor supervision[7].

To meet the requirements of detecting smoking behavior in small public places, we propose a YOLOv5[8] network based smoking behavior detection method. In addition, on the basis of researches of the above scholars and the YOLOv5s network model, a vertical rotation data enhancement method is employed to extend the dataset. The channel attention mechanism module is adopted to process the feature map obtained by convolution. Moreover, added a small target detection layer to the YOLOv5 algorithm to improve the target detection effect.

For clarity, our key contribution to this work can be summarized in three aspects:

1)We design a vertically-rotated data enhancement method to extend the dataset and thus increase the detected objects.

2)We introduce the SE-Net[9] (Squeeze-and-Excitation Networks) channel attention mechanism module to calibrate the feature response to improve the precision of detection .

3)We add a small target detection layer to the YOLOv5 algorithm,and to improve the detection accuracy

The remainder of the paper is structured as follows. The methods we adopt are described in detail in Section two.Section three describes the structure and components of the YOLOv5s network model. Section four demonstrates our experimental process and results. Section five serves as the summarization and conclusion of the paper.

2. Related works

2.1. Vertical rotation enhancement method

Every artificial intelligence system needs big data for training. In particular, artificial intelligence for object detection requires a lot of images for training. it has an impact on the training effect on the AI platform without proper pre-processing[10].

The YOLOv5s network model uses Mosaic data augmentation, merging four images through random scaling, cropping and arrangement. This paper employs the original enhancement method to add the process of data enhancement of the 90 degree vertical rotation in the pre-processing stage. Figure 1 shows the vertical rotation process, which named YOLOv5s-VR.

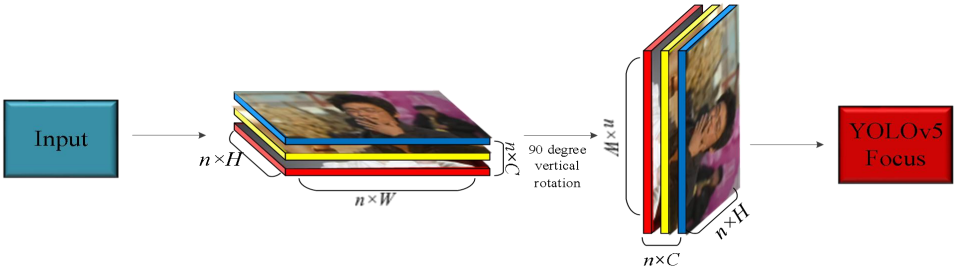

Figure 1. The work process of YOLOv5s-VR.

2.2. Attentional mechanism

Attention mechanisms have been widely applied in many areas of deep learning[11]. The core goal of attention mechanisms in deep learning is to select information which is more critical to the current task objectives from a large amount of information.

In computer vision, the mainstream attention mechanism can be classified into channel attention, spatial attention, and self-attention. The channel attention module reflects the relationship between different channels and feature graphs, automatically obtains the importance of each feature channel through network learning, and finally gives different weight coefficients to each channel. The spatial attention module essentially transforms the spatial information of the original image into another space while retaining the key information through the space conversion module, generating a weight mask for each position and weighting the output. The self-attention module is to reduce dependence on

external information and to interact as much attention as possible by using the information inherent in the feature.

Each channel represents a different semantic feature for feature graphs. The convolutional layer is primarily used to calculate the feature information in the feature graph, without considering the relationship between each feature channel[12]. This paper holds that the cigarette is a small object and has less feature information, thus requires attention to the feature information during the training process so that the trained model can better detect the cigarette object. Therefore, we introduce a channel attention module SE-Net into YOLOv5's neck network to improve the YOLOv5 network model.

The channel attention module can dynamically adapt to the completion to redefine the original features on the channel dimension, and focus on the dependencies at the model channel level. We set the SE-Net module after the backbone network stage of YOLOv5 after experimental verification. The SPP[13] module completes the spatial pyramid pooling operation and outputs three feature maps at the end of backbone network. The two dimensional features of each channel are compressed to one real number, and the number of channels remains the same. So it becomes $1 \times 1 \times C$ after the squeeze operation and obtains the global compressed encounter of the current feature graph. Then, through the excitation operation, the feature map is reduced to 1/r of the original feature dimension by FC (a fully connected layer). Secondly, it is activated by the ReLu function and generated back to the original feature dimension by FC. Then it is transformed into a normalized weight from 0 to 1 by the sigmoid function. Finally, it also inputs the weighted feature diagram into next layer's network. Figure 2 shows the overall structure of our improved model, which named YOLOv5s-SE.

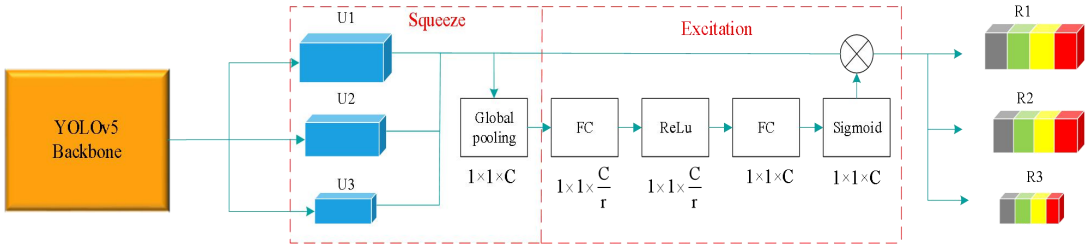

Figure 2. Structure of YOLOv5s-SE.

2.3. The small target detection layer

There are three detection layers in the original YOLOv5s model. An additional set of Anchor values is added in addition to the three initial Anchors values based on the original model as a way to detect smaller targets.

At the 17th layer of the neck network, operations such as upsampling of the feature map are performed so that the feature map continues to expand. Meanwhile, at the 20th layer, the feature maps obtained from the neck network are fused with the feature maps extracted by the backbone network. We add a small target detection layer of the prediction section at the 31th layer. In order to improve the detection accuracy, we use a total of four detection layers for the output feature map. As shown in Figure 3, we added a small target detection layer based on YOLOv5s algorithm, which named YOLOv5s-STD.

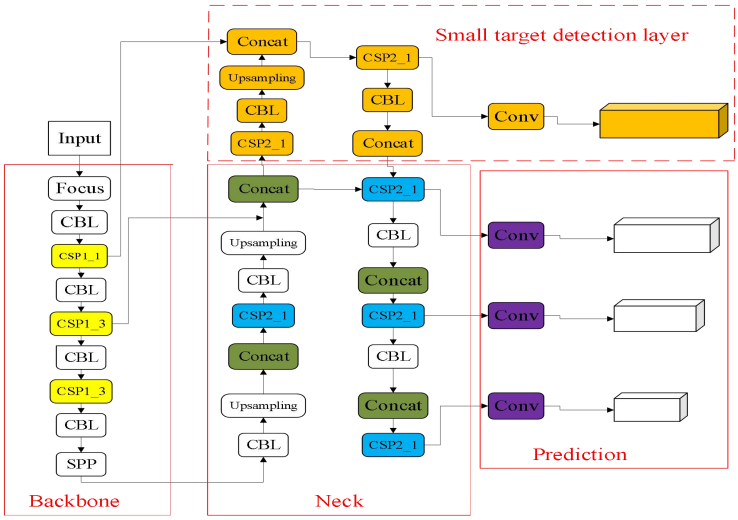

Figure 3. The overall structure of YOLOv5s-STD.

2.4. The work process of the improved algorithm

This paper is based on the YOLOv5 algorithm, which is improved in the input end, backbone network and prediction network. The specific workflow is shown in Figure 4, which named YOLOv5s-RST.

First, the image is input at the input end, and the vertical rotation operation is adopted for some images to increase the detection target. Secondly, feature extraction is performed on the image using convolutional neural network to generate feature maps. Then, the channels are weighted using the SE-Net module, and the weighted feature maps are output to the next network. Finally, the projection is completed by the improved prediction network.

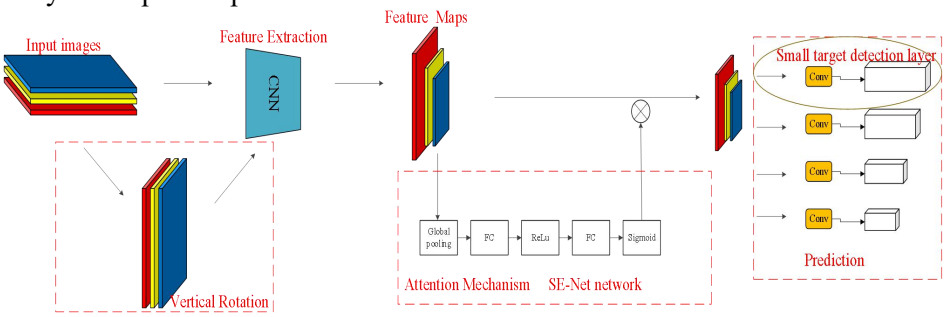

Figure 4. The work process of YOLOv5s-RST.

3. Network models introduction

With the great development of the artificial intelligence and computer vision technology over the past few decades[14],the models based on deep learning are used in smoking detection.The YOLOv4 model was released on April 23, 2020, followed by the release of YOLOv5 by Ultralytics LLC on June 10, 2020[15-16].YOLOv5 utilizes the Pytorch framework, and the weight file size of YOLOv5s with lightweight network structure is only 27MB[17]. Its speed for image inference is up to 140 frames per second, which can meet the needs of real-time detection.

The current mainstream of target recognition algorithms includes the R-CNN[18] series, SSD[19] and YOLO[20] series. The R-CNN series can be applied when target detection demands high precision, but its detection speed is slower than YOLOv5, and the need for real-time target detection cannot be satisfied in real application scenarios. SSD, though having slight advantages in speed and mAP over the YOLO series, has been surpassed by the newly proposed YOLOv5. Therefore, the YOLOv5 target detection algorithm is studied in this paper, and its YOLOv5s with lightweight network structure is shown in Figure 5. As can be seen from Figure 5, the network structure of YOLOv5s can be divided into four parts including input, backbone, neck and prediction.

The input of YOLOv5 uses mosaic's data enhancement method. YOLOv5 designs two CSP network structures, CSP1_x is used in the backbone part, and CSP2_x is used in the neck part. It can reduce the amount of calculation and ensure accuracy by using cross structure layer connections[21].

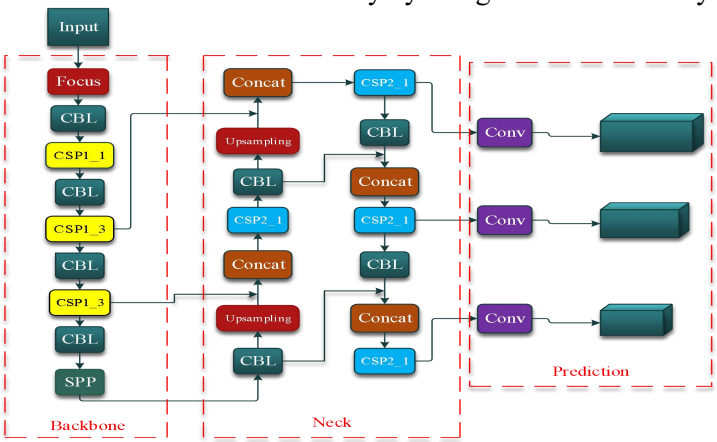

Figure 5. The YOLOv5s network structure.

4. Experimental analysis

4.1. Dataset creation and environment construction

The 12000 effective specimens were obtained by collecting online, intercepting smoking scenes in videos, and shooting smoking behavior in real life. The samples were divided into training sets and test sets in 0.7:0.3, with 8,400 samples in training sets, and 3,600 samples in test sets. And all valid samples were information tagged by the labelImg—a labeling software. The hardware and software indicators used in this experiment are shown in Table 1.

Table 1. Experimental platform and environmental configuration.

OS	CPU	GPU	Programming Lauguage
Windows10	Inter Core i5-10400F	GeFore GTX1650	Python3.7

4.2. Performance metrics

The model is tested by using the test set, The model's performance is evaluated by calculating the average accuracy of the model on the test set and the average accuracy average of all categories. The specific expressions are shown in (1)-(4).

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

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

$$AP = \int_0^1 P(R)dR \tag{3}$$

$$mAP = \frac{1}{|Q_R|}\sum_{q \in Q_R} AP(q) \tag{4}$$

Where FP means that the background is predicted as the number of targets while TP means that the number of matches between the predicted box and the true label box; FN represents that the number of targets to be detected but not detected by the model while QR refers to the number of test sets; AP is the average accuracy and mAP is the average accuracy of all categories.

4.3. Model training process

During the network training process, the loss function decreases with the number of iterations, which can reflect whether the network model converges or not. mAP is used to measure the superiority of the detection model. The higher the value, the better the model performance. The training process of the loss function value and mAP value of this model is shown in Figure 6.

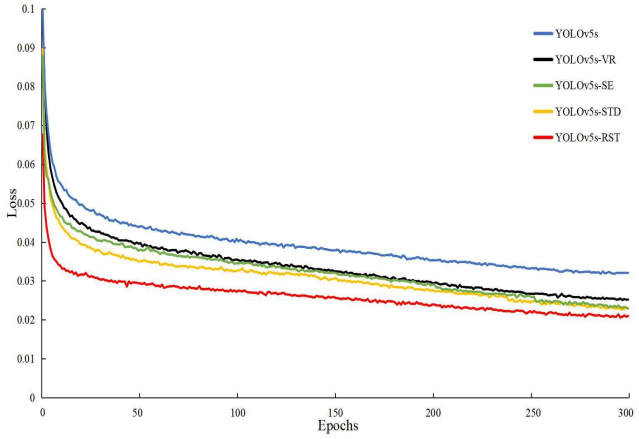

Figure 6(a). The training of loss.

As shown in Figure 6(a), each improvement work has better model performance than the original algorithm. As the number of iterations gradually increases, the loss value becomes smaller and the curve gradually converges. When the model is iterated 250 times, the loss value is basically stable and drops to around 0.02, and the network basically converges.

As shown in Figure 6(b), the YOLOv5s-RST model reaches 88.9% mAP after about 300 iterations and gradually stabilizes. The maximum value of the YOLOv5s-RST model reaches 90.1%, which is better than the YOLOv5s model.

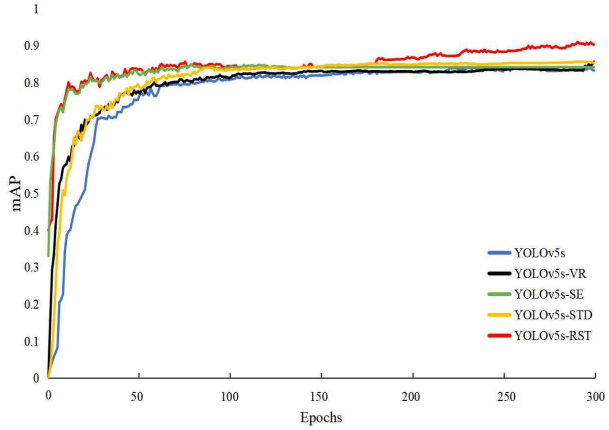

Figure 6(b). The training of mAP.

4.4. Analysis of experimental results

As an open-source deep learning framework by Facebook in 2017, Pytorch is selected as the experimental platform, and SGD optimization is adopted[22]. This study was conducted by comparative experiment and ablation experiment. The total iterations are 300, the initial learning rate is 0.01, and the weight attenuation coefficient is 0.0005, GIoU Loss is used as the loss function. The performance of the proposed method was compared with the original algorithm on a self-made smoking behavior dataset and the five model network performance metrics are shown in Table 2.

Table 2. Five model performance test comparisons.

Model	Precision(%)	Recall(%)	mAP(%)	Resolution ratio
YOLOv5s	87.9	79.7	83.6	640
YOLOv5s-VR	89.4	79.4	84.4	640
YOLOv5s-SE	89.4	80.7	84.6	640
YOLOv5s-STD	90.7	80.2	85.5	640
YOLOv5s-RST	90.3	82.4	88.9	640

According to Table 2, each of the improved work has better model performance than the original algorithm. The improved models have improved in terms of Precision, Recall and mAP values. The final improved model has improved mAP by nearly 5.3%.

Table 3. Comparison with YOLO series detection algorithms.

Model	YOLOv3	YOLOv3-tiny	YOLOv3-spp	YOLOv5s	YOLOv5s-RST
mAP(%)	79.2	72.5	75	83.6	88.9

It can be inferred from Table 3 that our network has a higher mAP compared to the YOLO series of detection algorithms. Our network has better performance to meet the detection needs in real life.

Figure 7 below shows the comparison between the detection results of the original algorithm and our network detection results in the actual smoking behavior detection.

(a) YOLOv5s.

(b) YOLOv5s-RST.

Figure 7. Comparison of detection results between the original algorithm and the improved algorithm.

As the results show, the improved network model can accurately detect the smoking behavior. For the same smoking behavior, the detection ability of the original algorithm is weaker. Under the condition that the IoU threshold is 45%, the mAP@0.5 of the original algorithm is 83.6%, and the mAP@0.5 of the finally improved algorithm is 88.9%.

5. Conclusion and future work

In this paper, deep learning technology was applied to smoking behavior detection and borrows ideas from the YOLOv5 algorithm. First, we add a vertical rotation data enhancement method to extend the dataset and increase the objects of detection. Secondly, we also introduce the channel attention module SE-Net at the end of backbone network to improve the feature extraction ability. Finally, we add a small target detection layer to the YOLOv5 algorithm, and to improve the detection accuracy. After the experiments, the proposed method can meet the demands of real-time detection, yet it still needs improvement in algorithm.

Subsequently, we will reduce the weight of the improved model and embed it into the mobile device. In the future, detection can be realized anytime, anywhere to meet the needs of daily use.

Funding Source

This work is in part supported by the General Project of Key R&D Program of Shaanxi Science and Technology Department (2017NY-129) and the General Special Project of Shaanxi Science and Technology Department (2021JQ-714).

Acknowledgments

We would like to thank Lei Cheng and Baidong Peng for providing English language support. We also thank Qiang Zhang and Chenbing Bai for their suggestions on the dataset.

References

[1] Ashare Rebecca L et al. The United States National Cancer Institute's Coordinated Research Effort on Tobacco Use as a Major Cause of Morbidity and Mortality among People with HIV[J]. Nicotine& Tobacco Research, 2021, 23(2), pp. 407-410 .

[2] Volkan Y.Senyurek et al. Smoking detection based on regularity analysis of hand to mouth gestures[J]. Biomedical Signal Processing and Control, 2019,51, pp.106-112.

[3] P. Wu, J. Hsieh, J. Cheng, S. Cheng and S. Tseng, "Human Smoking Event Detection Using Visual Interaction Clues," 2010 20th International Conference on Pattern Recognition, 2010, pp. 4344-4347.

[4] Chen Ruilong, Luo Lei, Cai Zhiping, et al. Real-Time Smoking Detection Algorithm based on Deep Learning [J]. Computer Science and Exploration, 2021,15 (02): 327-337.

[5] J. Liao and J. Zou, "Smoking target detection based on Yolo V3," 2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE), 2020, pp. 2241-2244.

[6] W. Cai, C. Wang, H. Huang and T. Wang, "A Real-Time Smoke Detection Model Based on YOLO-SMOKE Algorithm," 2020 Cross Strait Radio Science & Wireless Technology Conference (CSRSWTC), 2020, pp. 1-3.

[7] Z. Rentao, W. Mengyi, Z. Zilong, L. Ping and Z. Qingyu, "Indoor Smoking Behavior Detection Based on YOLOv3-tiny," 2019 Chinese Automation Congress (CAC), 2019, pp. 3477-3481.

[8] Yao Jia,Qi Jiaming,Zhang Jie,Shao Hongmin,Yang Jia,Li Xin. A Real-Time Detection Algorithm for Kiwifruit Defects Based on YOLOv5[J]. Electronics,2021,10(14).

[9] Jie, H.; Li, S.; Gang, S. Squeeze-and-Excitation Networks. IEEE Trans. Pattern Anal. Mach. Intell. 2017, 42, 7132–7141.

[10] H. Jeong, K. Park and Y. Ha, "Image Preprocessing for Efficient Training of YOLO Deep Learning Networks," 2018 IEEE International Conference on Big Data and Smart Computing (BigComp), 2018, pp. 635-637.

[11] Ren Huan, Wang Xuguang. Review of the attention mechanisms [J]. Computer Application, 2021,41 (S1): 1-6.
[12] X. Shi, J. Hu, X. Lei and S. Xu, "Detection of Flying Birds in Airport Monitoring Based on Improved YOLOv5," 2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP), 2021, pp. 1446-1451.
[13] He Kaiming,Zhang Xiangyu,Ren Shaoqing,Sun Jian. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition.[J]. IEEE transactions on pattern analysis and machine intelligence,2015,37(9).
[14] W. Zhang, L. Tian, C. Li and H. Li, "A SSD-based Crowded Pedestrian Detection Method," 2018 International Conference on Control, Automation and Information Sciences (ICCAIS), 2018, pp. 222-226.
[15] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, " Yolov4: Op-timal speed and accuracy of object detection," arXiv preprint arXiv: 2004.10934, 2020.
[16] Glenn J, Alex S, Jirka B, Nano Code012, Christopher STAN,Liu C.ultralytics: yolov5[J/OL], 2020 .3983579.
[17] Tan Shilei,Bie Xiongbo, Lu Gonglin, etc. Real-time detection of personnel masks based on YOLOv5 network model [J]. Laser Magazine, 2021,42 (02): 147-150.
[18] Yuan Huimin, Zhang Xuhong. Review of target detection algorithms [J]. Technology and Economics, 2021,29 (06): 52-55.
[19] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C.Berg, " Ssd: Single shot multibox detector," in European conference on computer vision. Springer, 2016, pp. 21-37.
[20] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition , 2016,pp.779-788.
[21] Z. Feng, L. Guo, D. Huang and R. Li, "Electrical Insulator Defects Detection Method Based on YOLOv5," 2021 IEEE 10th Data Driven Control and Learning Systems Conference (DDCLS), 2021, pp. 979-984.
[22] Zhang Jianlin, Ye Chunming, Li Zhaohui, et al. Analysis and prediction of the stock price by the Pytorch-based LSTM model [J]. Computer Technology and Development, 2021,31 (01): 161-167.