Small Sample Object Detection Based on Improved YOLOv5

Yuxuan Gao
School of Mechanical and Electronic Engineering, Beijing Jiaotong University
Beijing, Haidian District, China
Jiwu Wang
School of Mechanical and Electronic Engineering, Beijing Jiaotong University
Beijing, Haidian District, China
Zixin Li
Aero Engine Corporation of China,Beijing, Changping District, China
Email: 18815511536@163.com, jwwang@bjtu.edu.cn,15311426793@163.com
www.bjtu.edu.cn

Abstract

Object detection is widely used in various production and life, such as mask detection and recognition during the epidemic, face recognition with masks. Object detection algorithm based on deep learning has always been an important research content and implementation method in the field of object detection. Due to the large number of lead seals and fuses, their locations are not fixed, the lead seals and fuses have difficulties such as few sample datasets, complex target background and easy to be blocked, and strong reflective interference, and the conventional image processing methods are difficult to solve the problem of effective object recognition. In this study, by expanding the datasets, using different data enhancement methods, and training in the improved algorithm, the detection accuracy, detection speed, and adaptability were effectively improved.

Keywords: YOLOv5; Deep learning; Object detection; Few-shot learning

1. Introduction

Visual inspection, as a key application in related fields, provides fundamental support for achieving automation and is a key link in promoting the development of industrial automation [1]. Feature based visual detection methods have been widely used in industrial inspection, but this traditional detection method relies too much on manually designed feature extractors based on experience, and when faced with changes in target morphology and interference, the detection accuracy will significantly decrease. With the development of machine vision and artificial intelligence technologies, deep learning has become a new paradigm for object detection tasks in the industrial field [2].

However, deep learning relies on a large amount of labeled sample data during the training phase, which has drawbacks in real application scenarios. In factories, it is often difficult to obtain sufficient sample data due to issues such as a single model [3]. In such cases, the limited number of labeled samples can lead to severe overfitting of deep learning models and a decrease in recognition accuracy. Moreover, with sufficient datasets, labeling data can be time-consuming and labor-intensive, leading to issues of omission and mislabeling during labeling.

In this study, we used different image processing methods to further expand the original dataset and perform other data augmentations. At the same time, the automatic annotation of the dataset is completed through code to ensure the accuracy of the annotation, and it is trained in the improved YOLOv5 network to improve detection accuracy.

2. Methodology

2.1. Yolov5 algorithm

The core idea of YOLOv5 algorithm is to treat the object detection task as a regression problem and segment the feature map into L×L sized cells, M bounding boxes are set for each cell to surround the target. Each cell is responsible for detecting targets that fall within its area. In one operation, the bounding boxes, localization confidence, and probability of each category of targets contained in all cells are predicted. Then, the detected bounding boxes are filtered using NMS (Non Maximum Suppression) method to obtain the position, category, and confidence of the target to be tested.

The implementation of YOLOv5 algorithm can be summarized as the following steps:

(1) Preprocessing. The input image is first preprocessed, including scaling, cropping, and normalization operations, to convert it into a format suitable for network input.

(2) Feature extraction. The input image is fed into the backbone network and the feature is extracted through the backbone network.

(3) Feature pyramid network. YOLOv5 utilizes Feature Pyramid Networks (FPN) [4] to fuse feature maps of different scales, capture targets of different sizes, and improve detection accuracy.

(4) Predictive header. YOLOv5 uses a prediction head to predict each cell, obtaining bounding boxes (including the position and size of the target), confidence (probability of target existence), and class probabilities.

(5) Decoding prediction. Decode and convert the output of the prediction head into actual image coordinates, including adjusting the predicted bounding box to the original image size, and calculating the category and confidence score of each target.

(6) Non maximum suppression. Use NMS algorithm to remove redundant detection boxes with high overlap, and finally retain the most likely object detection result.

(7) Post processing. Perform post-processing on the detection results processed by NMS, including visualizing and saving the results.

The YOLOv5-7.0 network model structure is shown in Fig. 1, which mainly includes Conv, CSPDarkNet, and SPPF modules.

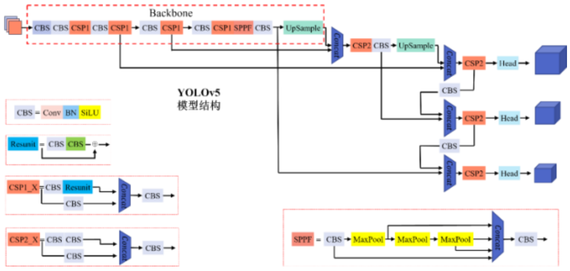


Fig.1 YOLOv5-7.0 model structure

2.2. SENetAttention

In order to obtain attention in the channel dimension, the input characteristic graph is convoluted to obtain the dimension H × W × C, and then compress the input characteristic graph through global average pooling. This step converts the characteristic graph of each channel into the scalar of 1 × 1 × C, thus forming a channel description vector, which captures the global spatial information of each channel and corresponds to the initial weight of each channel. This step is to compress (squeeze),as shown in equations (1) [5]:

$$Z_c = F_{sq}(x_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i,j) \qquad (1)$$

Where, H and W represent the height and width of the feature map, and x_c represents the input at each channel.

Then the second step is the excitation operation, which uses the small network of the Fully Connected (FC) layer to learn the relationship between channels. This network usually contains two full connected layers. The first full connected layer is responsible for mapping the channel description vector to a low dimensional space, with a ReLU activation function in the middle, and the second full connected layer maps it back to the original number of channels, This process can be regarded as a bottleneck structure, which aims to extract the dependencies between channels. The mathematical expression of this process is shown in Formula (2):

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \sigma(W_1 z)) \qquad (2)$$

Where, σ Represents the ReLU function, W_1 represents the parameters of the dimension reduction layer, W_2 represents the parameters of the dimension increase layer, and z is the output result from the previous compression step.

The last step is Re-scale, which multiplies the previously learned channel weight vector with the original feature map to achieve the function of adaptively adjusting the feature response of each channel and improve the performance of the network model. The network architecture of its algorithm module is shown in Fig. 2.

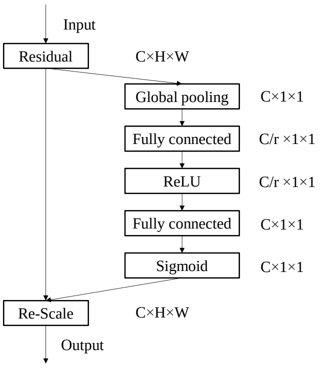


Fig.2 The diagram of the squeeze-and-excitation block

Adding the SENet attention module to the last layer of the Backbone network can effectively reduce the false detection rate and missing detection rate of lead seals and fuses, and improve the detection accuracy.

3. Data processing

3.1. image processing

For the image acquisition work of the product object, considering the complex surface structure of the product and the obvious reflective effect of the metal material and paint surface, the shooting effect is easily affected by the lighting conditions. According to this feature, the product is placed in a variety of lighting conditions, and different types of camera equipment are used to collect images. At the same time, local images and unmarked background images are taken in each lighting environment, in order to maximize the diversity of target images.Sample is shown in Fig. 3.

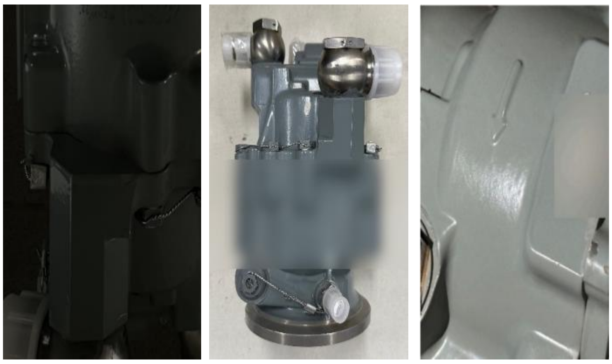

a) Low brightness environment b) Normal lighting environment c) High brightness environment

Fig.3 Shooting in various lighting conditions

ImgAug is a powerful and flexible image enhancement library based on Python language, which is mainly used in the field of deep learning and computer vision. It supports a variety of image processing libraries such as OpenCV. By calling ImgAug's API, the image can be transformed in various aspects such as geometry, color, contrast, etc., so as to expand the original image data set and improve the generalization ability and robustness of the model. With the help of its official API document, the corresponding Python code is written, and the image is transformed as follows. The process can be shown in Fig. 4.

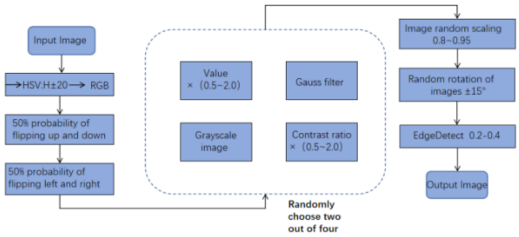
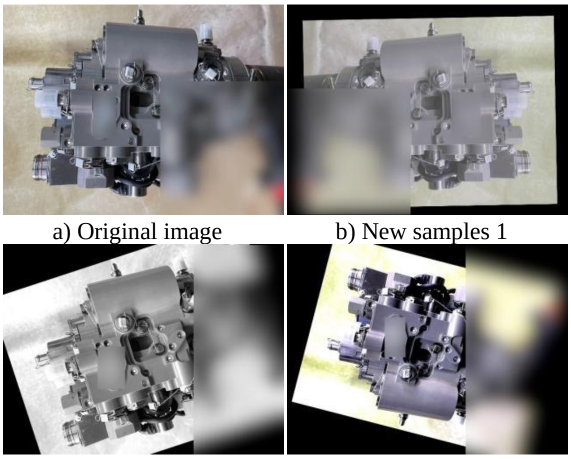

Input Image
→HSV.H±20→ RGB
50% probability of flipping up and down
50% probability of flipping left and right
Value × (0.5~2.0)
Gauss filter
Grayscale image
Contrast ratio × (0.5~2.0)
Randomly choose two out of four
Image random scaling 0.8~0.95
Random rotation of images ±15°
EdgeDetect 0.2-0.4
Output Image

Fig.4 Image enhancement process

Write python code to execute the above image enhancement operation, and realize the automatic annotation function of new images while realizing various transformations of images. By reading the XML annotation file corresponding to the original image, the position and coordinate information of all the annotation boxes are extracted, and according to the random enhancement process of each new image, the code automatically calculates and adjusts the information of the annotation box, thus generating a new XML file that is completely aligned with the target position in the new image. This method avoids the tedious work of manual marking and saves a lot of human resources. By automatically updating the annotation information, the enhanced image is consistent with the original annotation, thus improving the accuracy and reliability of the model. The new sample image of the original image after image enhancement is shown in Fig. 5.


a) Original image b) New samples 1

c) New samples 2 d) New samples 3

Fig.5 New samples after image augmentation

Through the above image enhancement method, each initial image is expanded to 6 new images, and finally 441 images are obtained as the final data set.

4. Experiment and Comparison

After completing the dataset production and modifying the YOLOv5 model, conduct training and compare the training results of the unimproved model. At the same time, verify the accuracy of the same image.

Baseline is the yolov5 model without adding SE attention mechanism, and the training is carried out using the non-expanded dataset. The dataset is as shown in the figure. SENet is the yolov5 model with adding SE attention mechanism, and the training is carried out using the non-expanded dataset. SENet+dataset is the yolov5 model with adding SE attention mechanism, and the training is carried out using the expanded dataset. The unexpanded dataset has 63 pictures, and the expanded dataset has 441 pictures, as shown in the following Fig. 6.

a) Baseline b) Expand the dataset

Fig.6 Image expansion

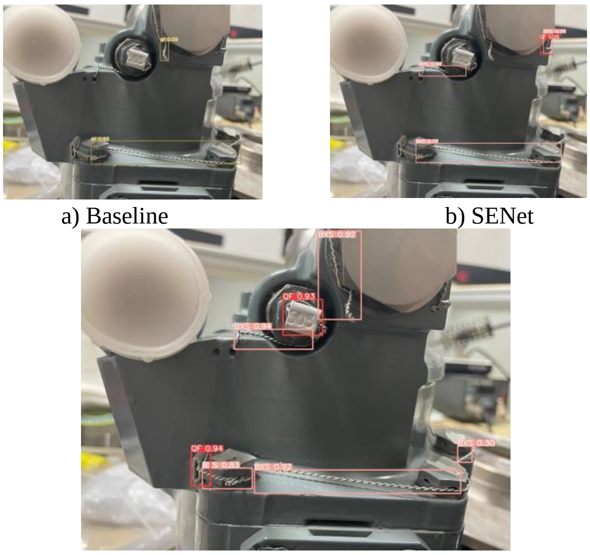
The key data of the training results are shown in Table 1. Among them, mAP (mean Average Precision) refers to the average precision, which comprehensively considers precision and recall to measure the performance of the model on detecting targets.

Table1. Comparison of data under different training conditions

	mAP (%)	precision (%)	time (ms)
Baseline	81.4	81.4	9.9
SENet	87.5	89.5	9.3
SENet+ dataset	96.3	96.9	12.3

From the table, it can be seen that compared to the baseline reference benchmark, the use of SENet attention mechanism in mAP has improved by 6.1%, and the combination of SENet and expanded dataset has improved by 8.2%; In terms of accuracy, compared to Baseline, the use of SENet attention mechanism improved by 7.9%, and the combination of SENet and expanded dataset improved by 7.4%.

The above data results demonstrate the advantages of SENet+augmented dataset. The best weight files trained using Baseline, SENet, and SENet+augmented dataset are tested on the images in the test set. The image detection results are shown in Fig. 7.



a) Baseline

b) SENet

c) NSENet+ Expand the dataset

Fig.7 Comparison of detection under different training conditions

From the figure, it can be seen that the effect is significantly improved after adding the SENet module network. However, due to the limited amount of data, there are still issues of false positives and missed alarms. After expanding the dataset, it can be clearly observed that not only is the accuracy further improved, but also false positives are reduced. In summary, the addition of SENet attention mechanism and the scalability of the dataset are helpful for model training and improving detection accuracy.

5. Conclusion

In this study, by improving the YOLOv5 model and expanding the dataset, the training accuracy is improved. The experimental results show that this method can meet the measurement accuracy requirements. Compared with the unimproved method, the method proposed in this paper improves detection accuracy and accuracy. This method can provide good support for deep learning based object detection research. However, due to the increased computational complexity of the algorithm, the running time is longer. In order to meet the real-time requirements of object detection, the speed of the algorithm needs to be further improved.

References

1. Fu Binbin Application and development trend of industrial machine vision [J] China Industry and Informatization, 2021 (11): 18-24
2. Guo Jing, Luo Hua, Zhang Tao Machine Vision and Application [J] Electronic Science and Technology, 2014, 27 (07): 185-188
3. Xing Z, Chen X. Lightweight algorithm of insulator identification applicable to electric power engineering[J]. Energy Reports, 2022, 8: 353-362.
4. Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.
5. Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

Authors Introduction

Dr. Jiwu Wang

He is an associate professor, Beijing Jiaotong University. His research interests are Intelligent Robot, Machine Vision, and Image Processing.

Mr. Yuxuan Gao

He is a postgraduate in Beijing Jiaotong University. His research interests are deep learning and yolo.

Ms. Zixin Li

She is an engineer at Aero Engine Corporation of China. Her research direction is 3D reconstruction