A Fabric Defect Detection System Based Improved YOLOv5 Detector

Ying Wang1, Zhengyang Hao2* Fang Zuo3 and Shanshan Pan4

1 Institute of Intelligence Networks System, Henan University, Kaifeng, Henan, 475001, China
2 Henan International Joint Laboratory of Theories and Key Technologies on Intelligence Networks, Henan University, Kaifeng, Henan, 475001, China
3 Intelligent Data Processing Engineering Research Center of Henan Province, Kaifeng, Henan University, Kaifeng, Henan, 475001, China
4 Henan International Joint Laboratory of Theories and Key Technologies on Intelligence Networks, Henan University, Kaifeng, Henan, 475001, China

*Corresponding author’s e-mail: 1510252872@vip.henu.edu.cn

Abstract. Fabric defect detection is a key part of product quality assessment in the textile industry. It is important to achieve fast, accurate and efficient detection of fabric defects to improve productivity in the textile industry. For the problems of irregular shapes and many small objects, an improved YOLOv5 object detection algorithm for fabric defects is propose. In order to improve the detection accuracy of small objects, the ASFF(Adaptively Spatial Feature Fusion) feature fusion method is adopted to improve the PANet's bad effect on multi-scale feature fusion. The transformer mechanisms can enhance fused features, allowing the network to focus on useful information. Experimental results show that the mean average precision of the improved YOLOv5 object detection algorithm in fabric defect map detection is 71.70%. The improved algorithm can quickly and accurately improve the accuracy of fabric defect detection and the accuracy of defect localization.

1. Introduction

Fabric quality has become an important factor affecting textile production efficiency in textile production. Therefore, the use of automatic defect detection technology has become an inevitable trend to improve fabric quality and reduce labor costs. The traditional methods of fabric defect detection are structure-based analysis, model-based analysis, and spectrum-based analysis[1], which are used frequently. KarleKar et al.[2] proposed a combined wavelet transform and morphology method to extract defect features by detecting fabric texture information. Jia et al.[3] automatically segmented the grid patterned fabric with repeating patterns by morphological processing and then detected the defect information by Gabor filter. LI Min et al.[4] proposed a new method with an improved Gaussian mixture model to detect defects in printed fabrics. However, the traditional algorithm has the problems of poor defect segmentation performance, noise sensitivity, and high miss detection rate.

Defect detection of fabric can be considered as an object detection problem. Deep learning models have powerful feature representation capabilities, and deep learning methods have a significant advantage over detectors with artificially set features in the object detection problem. These deep
learning methods accomplish two tasks when performing object detection: localization and classification of the detected frames. Deep learning-based object detection methods are mainly divided into two-stage object detection algorithms and one-stage object detection algorithms. The two-stage object detection algorithm implements the detection of objects in two stages: Firstly, the proposed regions are extracted from the images input to the network, then classification and location regression are performed for each proposed region. Finally, object detection is implemented. Such algorithms are typically represented by R-CNN[5], SPP-Net[6], Fast R-CNN[7] and Faster R-CNN[8]. At the same time, the one-stage object detection algorithm does not require a region proposal stage, but directly generates the class probability and position coordinate values of the object. Typical representatives of such algorithms are YOLO[9], SSD[10], YOLOv2[11], YOLOv3[12], RetinaNet[13], YOLOv4[14], YOLOv5[15].

The existing fabric detection algorithm has problems such as low efficiency of defect recognition and low recognition accuracy. Therefore, an improved YOLOv5 detection algorithm is proposed in this paper. The contributions to our work are as follows:

1. Aiming at the problem of irregular shape and the existence of more small objects, a Fabric defect object detection algorithm based on YOLOv5 improvement is proposed.
2. In this paper, we adopt the ASFF module to improve the PANet structure and fuse the feature maps of different layers by learning to get the weight parameters. It can filter features at other levels and keep the useful information at that level to improve the detection accuracy of small objects.
3. In this paper, we adopt the transformer mechanism to design the TransformerLayer module. It can increase the weight of useful features and suppress the weight of invalid features, which can improve the accuracy of object detection.

2. Related works

YOLOv5 is based on the original YOLO object detection architecture and uses the best optimization strategies in the field of convolutional neural networks in recent years, proposed by Glenn Jocher in 2020. The YOLOv5 backbone network combines the Focus structure and the CSP[16] structure to extract the main information in the input samples. Yolov5 uses the PANet structure to fuse the feature layers and achieves predictions on three different scales of the feature layers. The YOLOv5 architecture contains four architectures, specifically named YOLOv5s, YOLOv5m, YOLOv5l and YOLOv5x. The main difference among them is that the amount of feature extraction modules and convolution kernel in the specific location of the network is different. Therefore, the accuracy, efficiency and size of the recognition model were considered comprehensively in the study, and the improved design of the fabric object recognition network was carried out based on the YOLOv5s architecture. Figure 1 shows the network structure of YOLOv5s.

![Figure 1. The network structure of YOLOv5s.](image-url)
3. Materials and Methods

3.1. Adaptively Spatial Feature Fusion

YOLOv5s uses PANet structure to fuse and output multi-scale feature maps. The PANet is a bottom-up enhancement structure based on the FPN, which is a two-way fusion instead of the original single fusion. As shown in Figure 2, PANet is composed of top-down fusion path (a) and bottom-up fusion path (b). The top-down fusion is shown in Figure 3. It adapts the nearest feature map X by 2-fold up-sampling and adds it to the pre-layer feature map Y by 1×1 convolution after adjusting the channels. Bottom-up fusion is the opposite process of top-down fusion, where up sampling is replaced by down-sampling.

![Figure 2. PANet structure.](image)

![Figure 3. Top-down integration.](image)

This fusion method just converts the feature maps to the same size and then adds them together, which does not take full advantage of the features at different scales. In this paper, we introduce a new spatial fusion approach, adaptive spatial feature fusion (ASFF)[17], to improve the PANet structure. It fuses the feature maps of different classes by learning to get the weight parameters. The Improved PANet structure is shown in Figure 4.

![Figure 4. The Improved PANet structure.](image)

X1, X2, X3 respectively represent the feature maps extracted by the backbone network. For example, ASFF3 get feature maps Level1 and Level2 after PANet structure. Level1 and Level2 are compressed into the same number of channels as Level3 by 1×1 convolution and respectively up-sampling by 4 times and 2 times to form feature maps of the same dimension as Level3. The generated feature map is denoted as resize_Level1 and resize_Level2. Then resize_Level1, resize_Level2 and Level3 are convolved by 1×1 to get the weight parameters α, β and γ. Finally, resize_Level1, resize_Level2 and Level3 are respectively multiplied by A, B and C and added to get the new fusion features.

The ASFF structure can filter features at other levels and keep the useful information at that level to improve the detection accuracy of small objects. Equation (1) shows the ASFF calculation function.

\[ y'_i = \alpha'_i \cdot x^{1,i} + \beta'_i \cdot x^{2,i} + \gamma'_i \cdot x^{3,i} \quad (1) \]
In the ASFF calculation function, \( y'_j \) represents the new feature graph obtained by ASFF. \( x^{a-m}_{ij} \) represents the feature vector from layer \( n \) to layer \( l \) on the feature map. \( \alpha'_{ij}, \beta'_{ij}, \gamma'_{ij} \) represent the weight values of three different levels of feature maps, which are made to satisfy \( \alpha'_{ij} + \beta'_{ij} + \gamma'_{ij} = 1(\alpha'_{ij}, \beta'_{ij}, \gamma'_{ij} \in [0,1]) \) by softmax function.

3.2. The transformer mechanism
Transformers were first proposed by [18] for machine translation and established state-of-the-arts in many NLP tasks. To make Transformers also applicable for computer vision tasks, several modifications have been made. For instance, Parmar et al. [19] applied the self-attention only in local neighborhoods for each query pixel instead of globally. Child et al. [20] proposed Sparse Transformers, which employ scalable approximations to global self-attention. Recently, Vision Transformer (ViT) [21] achieved state-of-the-art on ImageNet classification by directly applying Transformers with global self-attention to full-sized images. We adopt the core idea of Transformer and exploit the attention mechanism to design the TransformerLayer module. The structure of TransformerLayer is shown in Figure 5. The TransformerLayer contains the Multi-Head Attention structure. The structure of Multi-Head Attention is shown in Figure 6.

\[
\text{MultiHead}(Q, K, V) = \text{Concat}(H_1, \cdots, H_{n_h})W^O
\]

\[
H_i = \text{Attention}(QW_i^O, KW_i^K, KW_i^V)
\]

where \( W_i^O \in \mathbb{R}^{d_{x} \times d_{x}}, W_i^K \in \mathbb{R}^{d_{x} \times d_{x}}, W_i^V \in \mathbb{R}^{d_{v} \times d_{v}}, \) and \( W^O \in \mathbb{R}^{d_{x} \times d_{x}} \) are parameter matrices.

\[
\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V
\]
The TransformerLayer module is based on multi-head self-attention in a residual form. The TransformerLayer is defined in equations (5)(6).

\[ X_{output} = X_{output}^* + (\hat{X}_{output}) W^2 W \]  
\[ \hat{X}_{output} = X + \text{MultiHead}(Q, K, V) \]

3.3. The improved network structure

In this paper, we adopt the ASFF module to improve the PANet structure and fuse the feature maps of different layers by learning to get the weight parameters. It can filter features at other levels and keep the useful information at that level to improve the detection accuracy of small objects. Then, we adopt the transformer mechanism to design the TransformerLayer module. It can increase the weight of useful features and suppress the weight of invalid features, which can improve the accuracy of object detection. Figure 7 shows the improved network structure of YOLOv5s.

4. Results and Discussion

4.1. The fabric dataset

The dataset used in this paper is from a textile workshop in Foshan City, Guangdong Province. The dataset contains 3946 defective images with a resolution of 1000×1000. The object size of the dataset is mostly small. Figure 8 shows the samples of fabric defects. Figure 9 shows the shape distribution of fabric defects. Table 1 shows the sample distribution of the dataset.
Figure 8. The samples of fabric defects.

Figure 9. The shape distribution of fabric defects.

Table 1. The sample distribution of the dataset.

| Category       | Hole  | Heterosexual fiber | Knot  | Broken spandex | Total |
|----------------|-------|--------------------|-------|----------------|-------|
| Number         | 548   | 1762               | 2444  | 1044           | 5798  |
4.2. Evaluation metrics
In this paper, mAP50 is used as an evaluation metric in the experiments. mAP50 is the mean value of AP50 (average precision) of the detection results for all categories. The AP50 value is the enclosed area of the accuracy and recall curves for an IoU threshold of 0.5. Equation (7) (8) show the calculation of precision and recall values. Equation (9) (10) show the calculation of AP50 and mAP50.

\[
\text{precision} = \frac{TP}{TP + FP} \quad (7)
\]

\[
\text{recall} = \frac{TP}{TP + FN} \quad (8)
\]

TP (True Positives) indicates the number of objects actually detected in the dataset. FP(False Positives) indicates the number of objects detected by the detection model in error. FN(False Negatives) indicates the number of objects missed by the detection model.

\[
AP50 = \int_0^1 \text{P}(r, \text{IoU} \geq 0.5) \quad (9)
\]

\[
mAP50 = \frac{\sum_{i=1}^{n} AP50_i}{N} \quad (10)
\]

4.3. Experimental Environment and Configuration
In this paper, 80% of the training set are used for training and 20% for verification. The GPU is NVIDIA GeForce P100 and the experimental environment is Pytorch. The input size is 608×608×3. In this study, the improved YOLOv5s network was trained by stochastic gradient descent (SGD) in an end-to-end way. The batch size of the model training was set to 8, and each time, the regularization was done by the BN layer to update the weight of model. The momentum factor (momentum) was set to 0.937, and the decay rate (decay) of weight was set to 0.0005. The number of training epochs was set to 100. The learning rate was set to 0.013.

4.4. Test result and Analysis
The algorithm proposed in this paper is based on YOLOv5s improvement. In order to compare the magnitude of the performance difference between the two algorithms, the metric changes during the training process are printed out separately for comparison. The experimental results are shown in Figure 10. The red curve is the improved Yolov5s and the gray curve is the original Yolov5s.

![Figure 10](image-url) This algorithm is compared with YOLOv5s.

As can be seen in Figure 9, the algorithm in this paper achieves better performance in both mAP values than the original YOLOv5s algorithm.
In order to evaluate the algorithm performance, the improved YOLOv5s proposed in this paper were compared with YOLOv3, YOLOv4, YOLOv5s, Faster R-CNN and RetinaNet in various aspects. The algorithm comparison experiments were performed using the same training and test sets, and the AP50 and mAP50 results of the different models on the datasets are shown in Table 2.

**Table 2. Different network detection results.**

| Model            | Input size | AP50  | mAP50  |
|------------------|------------|-------|--------|
| Faster R-CNN     | 608*608    | 82.30%| 65.20% |
| RetinaNet        | 608*608    | 84.30%| 56.40% |
| YOLOv3           | 608*608    | 71.60%| 68.80% |
| YOLOv4           | 608*608    | 84.10%| 65.60% |
| YOLOv5s          | 608*608    | 84.70%| 75.40% |
| Improved YOLOv5s | 608*608    | 86.80%| 79.80% |

In order to reflect the performance of the improved object detection algorithm more intuitively. The YOLOv5s-based object detection model (the original algorithm) is firstly used for detection, and then the improved YOLOv5s algorithm model is put into detection. Figure 11 shows the detection results of YOLOv5 and improved YOLOv5s. Columns (a) show the detection results of the YOLOv5s algorithm. Columns (b) show the detection results of the improved YOLOv5s algorithm. In the detection picture, label 1 represents the hole, label 3 represents the Heterosexual fiber, label 4 represents the knot and label 17 represents the broken spandex. From the experimental results, it is obvious that YOLOv5s has some missed detection objects and the confidence level is very low. The improved YOLOv5s algorithm in test images with good detection and high confidence. In summary, the improved YOLOv5s algorithm proposed in this can comply with the requirements of real-time detection while maintaining a high detection rate.
5. Conclusions
In this paper, these improvements based on the YOLOv5s algorithm make the object detection with higher accuracy and detection speed. In the future, we will consider designing our own network structure to build a lightweight network with superior performance, so as to improve the detection accuracy and real-time performance of the model.

Acknowledgments
We thank Henan International Joint Laboratory of Theories and Key Technologies on Intelligence Networks and Institute of Intelligent Network System of Henan University for providing experiment environment. This work is sponsored by Natural Science Foundation of Henan Province (No. 182300410164), Key Scientific Research Project of Universities in Henan Province (No. 19A520016), Henan Higher Education Teaching Reform Project and Practice Project (Graduate Education) (No. 2019SJGLX080Y), Henan University Graduate Education Innovation and Quality Improvement Plan Project (No. SYL19060120), Science and Technology Research Key Project of Education Department of Henan Province (No. 14B520003 and 16B520005).
References

[1] Yapi, Daniel., Allili., Mohand Said., Baaziz, Nadia. (2017) Automatic Fabric Defect Detection Using Learning-Based Local Textural Distributions in the Contourlet Domain. IEEE Transactions on Automation Science and Engineering., 1–13

[2] Vaibhav, M., Karlekar, V., Bhangale, K., Bira Da R, M. S. (2015) Fabric Defect Detection Using Wavelet Filter. In: Computing Communication Control and Automation. Pune. pp. 712–715.

[3] Jia, L., Chen, C., Liang, J., Hou, Z. (2017) Fabric defect inspection based on lattice segmentation and Gabor filtering. Neurocomputing., 238: 84–102

[4] Li, M., Cui, S., Xie, Z. (2015) Application of gaussian mixture model on defect detection of print fabric. Journal of Textile Research., 36: 94–98

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

[6] He, K., Zhang, X., Ren, S., Sun, J. (2015) Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE transactions on pattern analysis and machine intelligence., 37: 1904–1916

[7] Girshick, R. (2015) Fast r-cnn. In: Computer Vision and Pattern Recognition. Boston. pp. 1440-1448.

[8] Ren, S., He, K., Girshick, R., Sun, J. (2016) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence., 39: 1137–1149

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

[10] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C. (2016) SSD: Single Shot MultiBox Detector. In: European conference on computer vision. Amsterdam. pp. 21–37.

[11] Redmon, J., Farhadi, A. (2017) Yolo9000: better, faster, stronger. In: Conference on Computer Vision and Pattern Recognition. Hawaii. pp. 7263–7271.

[12] Redmon, J., Farhadi, A. (2018) Yolov3: An incremental improvement. arXiv:1804.02767

[13] Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P. (2017) Focal loss for dense object detection. In: International Conference on Computer Vision. Venice. pp. 2980-2988.

[14] Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M. (2020) Yolov4: Optimal speed and accuracy of object detection. arXiv:2004.10934

[15] Glenn Jocher Spain. (2020) Yolov5 code. https://github.com/ultralytics/yolov5.

[16] Wang, C. Y., Liao, H., Yeh, I. H., Wu, Y. H., Chen, P. Y., Hsieh, J. W. (2019). Cspnet: a new backbone that can enhance learning capability of cnn.

[17] Liu, S., Huang, D., Wang, Y. (2019). Learning spatial fusion for single-shot object detection.

[18] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I. (2017) Attention is all you need. In: Advances in neural information processing systems. Los Angeles. pp. 5998–6008

[19] Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., Tran, D. (2018) Image transformer. In: International Conference on Machine Learning. Stockholm. pp. 4055–4064.

[20] Child, R., Gray, S., Radford, A., Sutskever, I. (2019) Generating long sequences with sparse transformers. arXiv:1904.10509

[21] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2021) An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations.