Automation and Machine Learning (2025)
Clausius Scientific Press, Canada

DOI: 10.23977/autml.2025.060118
ISSN 2516-5003 Vol. 6 Num. 1

Fabric defect detection algorithm based on improved RT-DETR

Ruiming Liu[1,a], Shuai Huang[1,b,*], Xuesong Duan[1,c], Yunliang Du[1,d]

[1]School of Electronic Engineering, Jiangsu Ocean University, Lianyungang, Jiangsu, China
[a]liurm@jou.edu.cn, [b]2023220626@jou.edu.cn, [c]2023210601@jou.edu.cn,
[d]2023220622@jou.edu.cn
**Corresponding author*

Keywords: RT-DETR, fabric defect, feature fusion, attention mechanism, ablation experiment

Abstract: Textiles are important raw materials in industry and life, and China's textile industry plays a key role, but there are more than 80 kinds of surface defects in fabric production, which affect the quality and development of the industry. Current detection algorithms have problems such as insufficient accuracy and limited application scenarios. Manual detection and traditional machine vision methods also have obvious defects. Although algorithms based on deep learning have applications, they have their own shortcomings. Therefore, an improved RT-DETR fabric detection algorithm RT-FDTR is proposed in this study: optimizing the backbone network, introducing C2f_AdditiveBlock module to enhance feature extraction ability; designing DHSA-AIFI module to enhance small target detection and anti-interference ability; developing SCOK-CCFF feature pyramid to optimize feature fusion. Experiments on the fabric defect dataset of Aliyun Tianchi show that the P, R and AP50 of the improved model are 82.2%, 77.1% and 76.5% respectively, which are 6.9%, 2.5% and 3% higher than those of the original RT-DETR-r18, and the parameters are reduced by 20.6%. The detection speed is increased by 9.7FPS, which meets the accuracy and real-time requirements of fabric defect detection in industry.

1. Introduction

Textile is a kind of raw material widely used in clothing in industrial and living scenes, and has irreplaceable social value. China ranks first in the world in terms of textile production and export. Textile industry is one of the pillar industries of China's national economy and plays a key role in national development. There are more than 80 kinds of surface defects in the production process of cloth, which affect the quality of textiles and thus affect sales, limiting the development of textile industry[1-2], so accurate detection of cloth surface defects is indispensable[3]. The accuracy of current fabric defect detection algorithms cannot meet the industrial requirements, and most algorithms are only suitable for specific types of fabric defect detection. However, in practical industrial applications, fabric defects vary in size and shape, and existing detection algorithms are not suitable for such complex scenarios.

Early fabric defect detection relied on the human eye to complete. This method is not only inefficient, but also has high labor cost, and often misses and misjudges in the detection process[4]. Nowadays, there are new developments in detection methods, one of which is based on traditional machine vision, including model analysis, spectral analysis and statistical analysis. Among them, model analysis method constructs texture in image into random or deterministic model, which can be used to deal with random changes of fabric background texture. However, this method is extremely complex and computationally intensive, which is not suitable for detecting small targets. Spectral analysis is used to extract periodicity of texture primitives from spectral features of images. However, defects on complex surfaces cannot be detected; statistical analysis methods use the statistical characteristics of the relationship between image gray levels to extract the features of defective fabrics. When used alone, they require prior knowledge and are only effective for a few types of defects. These traditional methods need to design different feature extraction algorithms for different types of defects, which are not suitable for practical industrial applications.

Another category is target detection algorithms based on deep learning. With the rapid development of deep learning, it has been applied to a variety of scenarios such as image classification, object detection and fault diagnosis. Defect detection methods based on deep learning have been applied to various industrial products, especially liquid crystal panels, metal materials and textiles. In recent years, many scholars have proposed feasible textile testing methods. For example, Liu[5] et al. proposed a fabric defect detection method based on context-aware attention cascade feedback network, and designed a parallel context extractor to capture multi-scale context information to achieve accurate defect location. Chen[6] et al. used label smoothing strategy in training phase and data enhancement technique to compensate for imbalance of data set, reducing the number of model parameters without reducing model accuracy, and improving average detection accuracy by 1.3%. Zhang[7] et al. add coordinate attention mechanism to YOLOv8n model based on YOLOv8 algorithm. The improvement enhances the network's ability to extract fabric defect features. On DAGM10 dataset, mAP reaches 79.20%, detection speed reaches 95.4FPS, and can meet the requirements of online real-time detection with low computing resource requirements. It can be deployed in edge intelligent computing equipment, which promotes the application of textile defect automatic detection technology in industry. Hülya Gökalp Clarke[8] et al. constructed a convolutional neural network (CNN) for fabric defect detection, employing a cyclic learning rate scheduler and trained and validated using an original dataset containing three fabric types and four types of defects. This model can effectively identify many kinds of fabric defects. Li[9] proposed a fabric defect detection method combining hybrid attention transformer (HAT) and improved cascaded R-CNN (SPCNet), which effectively improved the detection performance of multi-class fabric defects. Zhang[10] et al. proposed several methods for color fabric detection, which used attention-based feature fusion to generate adversarial network and quadtree attention-based U-shaped Swin Transformer network to detect fabric defects respectively. By introducing some attention modules to enhance the extraction of defect features, and using U-shaped network to realize pixel-level reconstruction of images, the detection and location accuracy were improved. However, this method is mainly implemented through unsupervised detection by CNN. This learning method relies on random initialization and data distribution, resulting in the learned features being inaccurate or biased, and requiring more computing resources and time.

To solve these problems, the RT-DETR algorithm is faster, more accurate and more real-time than other mainstream algorithms. In this paper, a fabric detection algorithm based on improved RT-DETR is proposed. The experimental results show that the speed and accuracy of the whole model meet the industrial requirements.

2. Improved RT-DETR fabric defect detection method

As the first real-time end-to-end target detector, RT-DETR consists of backbone network, hybrid encoder, Transformer decoder and auxiliary prediction header[11]. Backbone network extracts multi-scale features through CNN, providing low-level, middle-level and high-level feature representations for subsequent processing; hybrid encoder consists of attention-based (adaptive feature integration, AIFI) module and CNN-based CCFM module, AIFI module enhances the hierarchy and richness of advanced features, CCFM (cross-level cross-feature fusion module) facilitates fusion and interaction of different levels of features; processed features generate target predictions through Transformer decoder, which assist prediction head to further optimize detection performance. RT-DETR improves target query initialization through IOU aware queries and can adjust decoder layers to optimize real-time performance.

In order to cope with the complexity of the environment and the limitation of the model size in fabric defect detection, a lightweight enhancement model RT-FDTR is proposed based on the real-time detection transformer with resnet-18 backbone (RT-DETR-r18). The structural design of the model is shown in Figure 1. Firstly, additive similarity function (CATM) is introduced into the backbone network to enhance the multi-scale feature extraction ability of the model in fabric defect detection. In complex scenes of cloth defects and cloth background textures, the recognition ability of defect targets is improved. This not only provides richer and more accurate feature information for subsequent modules, but also effectively reduces background interference and enhances the discrimination between target and background. On this basis, DHSA-AIFI module further strengthens the feature expression extracted by AdditiveBlock module by using multi-scale multi-head self-attention mechanism. DHSA-AIFI's multi-scale feature ensures that the model can capture the features of different scale targets when the defect target size varies greatly, and improves the discrimination ability between target and background by enhancing the learning of context information. DHSA-AIFI thus provides the model with an enhanced representation of detail information in complex background environments, enabling the model to better adapt to target defects at different scales, in different backgrounds and under different lighting conditions. Finally, SCOK-CCFF feature pyramid structure optimizes the multi-scale information extracted by backbone network and the detailed features aggregated by DHSA-AIFI by efficiently fusing shallow and deep features. SCOK-CCFF not only improves the efficiency of multi-scale information utilization, but also makes the model more robust in multi-scale target detection, especially for small targets. SCOK-CCFF can effectively enhance the discrimination between object and background in complex environment with low contrast between cloth defects and background texture, so that the model can accurately detect defects in complex background and improve the robustness of detection.

To sum up, the backbone network provides enhanced multi-scale features, DHSA-AIFI module enhances feature expression through multi-scale self-attention mechanism, and SCOK-CCFF feature pyramid structure further improves the robustness of the model and the recognition ability of small targets by optimizing the fusion efficiency of features. The synergy of these modules ensures that the RT-FDTR model can more accurately identify and locate defects in complex environments, especially in dealing with small targets and background disturbances, improving the overall detection accuracy and robustness of the model. The following sections detail the specific design details of the enhanced model.

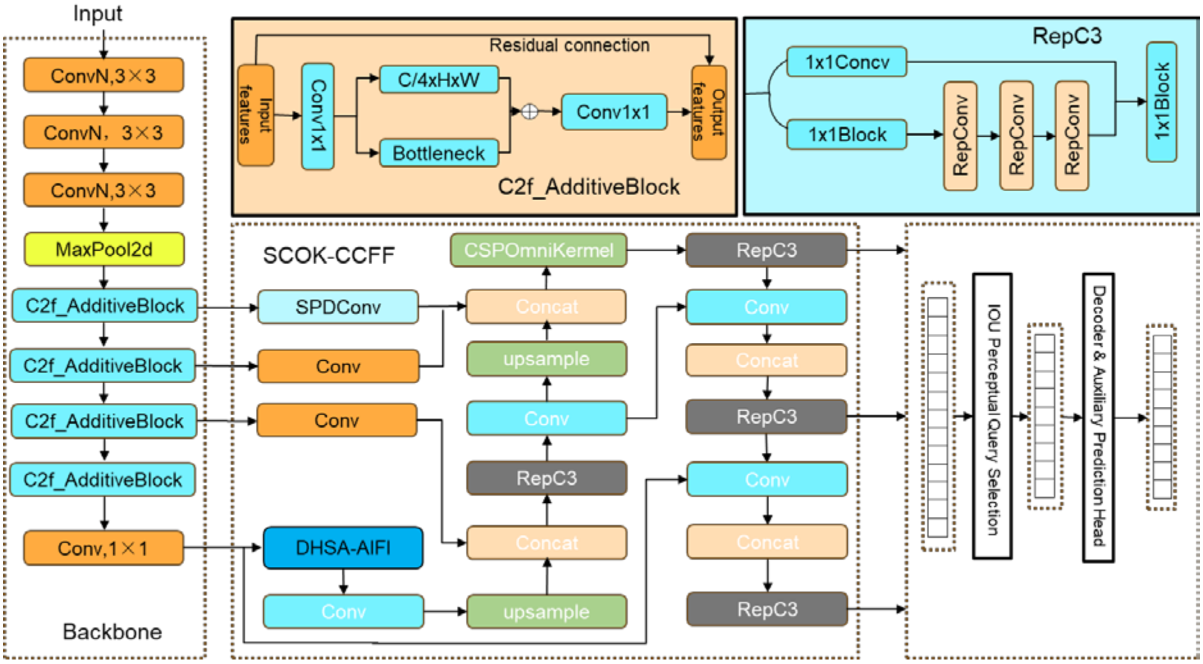


Figure 1: Network structure diagram.

2.1. Improvement of backbone network

RT-DETR's backbone network is based on a (residual network, ResNet) architecture, which uses a four-layer (basic residual block, BasicBlock) structure to balance computational efficiency and performance[12]. However, BasicBlock has limitations in cloth defect detection tasks: small receptive fields limit the capture of global context information, making it difficult to process global information of large targets and local details of small targets simultaneously; lack of multi-scale feature fusion leads to a decline in detection accuracy under complex backgrounds; in addition, higher parameters increase the computational burden and affect the efficiency of model deployment.

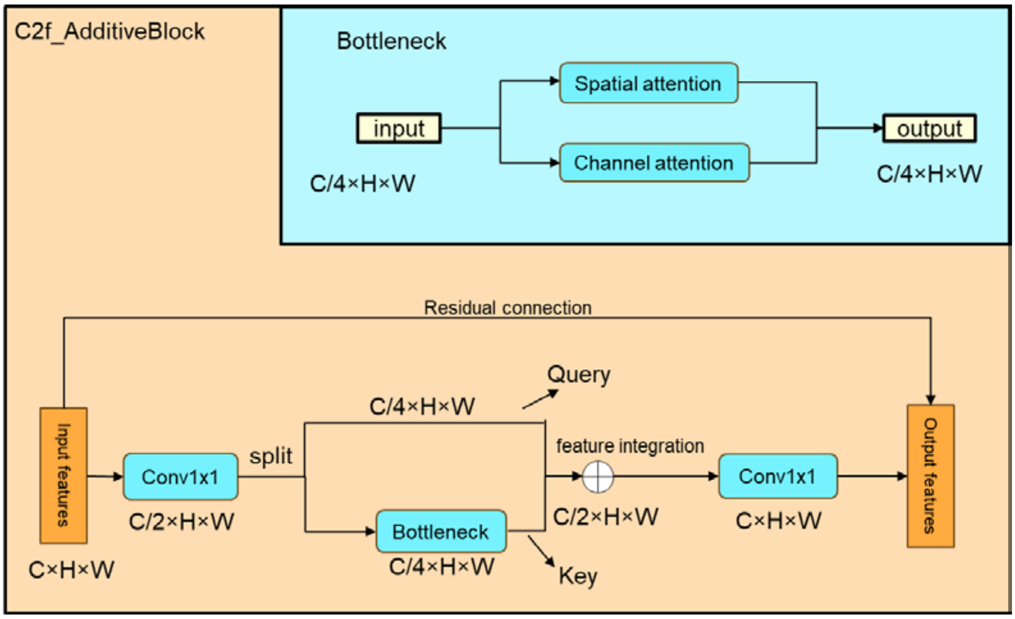


Figure 2: Structure diagram of C2f_AdditiveBlock module.

In order to solve these problems, this paper proposes a C2f_AdditiveBlock module based on CAS-ViT (Convolutional Additive Self-attention Vision Transformers). The module reduces the dimensionality by 1×1 convolution, and then splits the dimensionality reduced features into two branches according to 1:1. Branch 1 is directly transferred across stages, retaining the low-frequency basic features of the original fabric texture. Branch 2 enters the spatial-channel dual-domain attention interaction process and is processed by Bottleneck module. Bottleneck is divided into spatial attention branch and channel attention branch for dual-domain feature fusion. Meanwhile, SENet integrates channel information by 1×1 convolution. The similarity function is innovatively defined as the sum of the context scores of Query and Key, avoiding complex operations such as matrix multiplication and Softmax, and reducing computational complexity. Secondly, referring to the convolutional additive self-attention (CAS) block hybrid architecture, the input features and output features are connected by residuals, so that the underlying network is realized by convolution, and a good balance is achieved between computational efficiency and deployment.

Figure 2 shows the structure diagram of the C2f_AdditiveBlock module. The module inputs the size of the feature diagram as C×H×W, and input $x \in R^{H \times W \times C}$. Firstly, the dimension of input features is reduced to C/2×H×W by 1×1 convolution layer, which reduces the amount of subsequent calculations. Then, the dimensionality reduced features are divided into two branches along the channel dimension. Branch 1 directly transfers the split features across stages and waits for feature fusion with the output features of branch 2. Finally, the original dimension is restored through 1×1 convolution layer, and the input features and output features are connected through residuals.

Bottleneck splits the input into spatial domain attention branches and channel domain attention branches in branch 2, followed by feature fusion through the operation of "stacking." The spatial attention structure diagram is shown in Figure 3. First, the receptive field is enlarged by a 3×3 convolutional layer, then the channel is compressed by a 1×1 convolutional layer, and then the Sigmoid activation function is used to generate the spatial attention weight map. The spatial attention branch is specifically expressed as:

$$x_s = Sigmoid(D_{1\times1}(D_{3\times3}, \mathrm{Re}\,LU, BN(x))) \odot x \qquad (1)$$

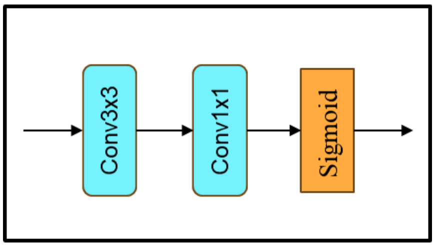


Figure 3: Spatial attention branches.

For channel-domain attention, we refer to SENet, which, as shown in Figure 4, does not use channel reduction, but instead uses 1×1 convolution to integrate information between channels:

$$x_c = Sigmoid(D_{1\times1}(P_{1\times1}(x))) \odot x \qquad (2)$$

Where P represents adaptive pooling, D is a grouped convolutional layer, and the default number of packets is set to the number of channels. The Bottleneck module in Branch 2 stacking these two operations results in a feature map of the interaction between space and channel domains, denoted as $\Phi(x)$.

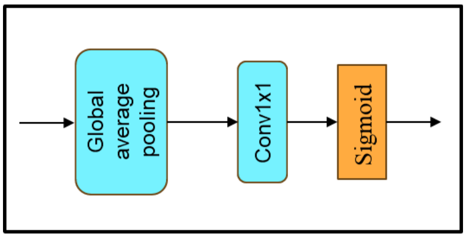


Figure 4: Channel domain attention branch.

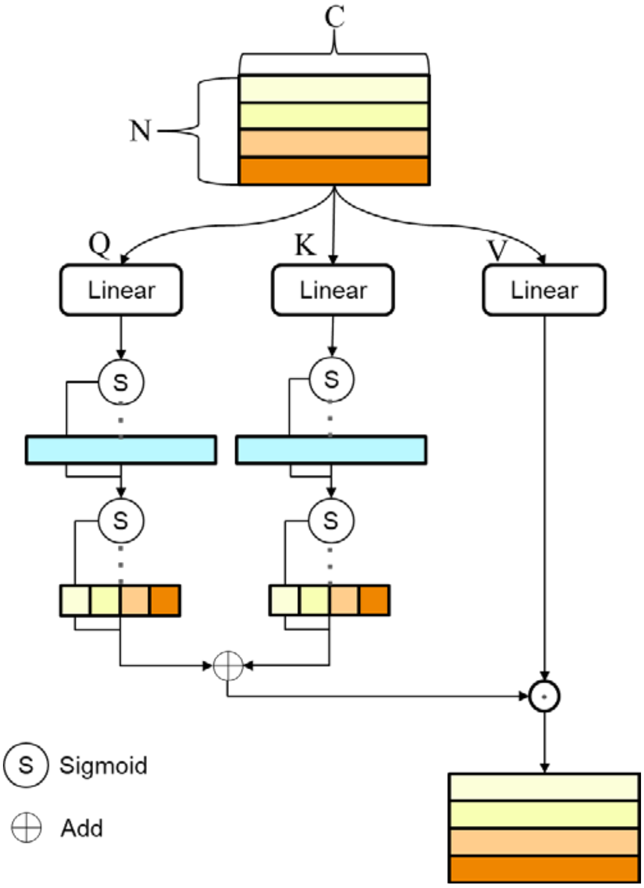


Figure 5: Conv Additive Self-Attention.

Finally, we define the similarity function as the sum of the context scores of Q(Query) and K(Key), i.e. branch 1 cross-stage features and branch 2, adding fused features as shown in Figure 5:

$$Sim(Q,K)=\Phi(Q)+\Phi(K) \tag{3}$$

Query, Key and Value are obtained through independent linear transformation, $Q=W_q x$, $K=W_k x$, where $V=W_v x$, W_q, W_k and W_v are learnable weights, $\Phi(\bullet)$ is a context mapping function, which contains necessary interactive information, and the complexity can be reduced by adding methods while retaining effective information. Finally, the output can be expressed as:

$$O=\Gamma(\Phi(Q)+\Phi(K))\odot V \tag{4}$$

Where $\Gamma(\bullet)\in R^{N\times C}$ represents a linear transformation of the set of context information.

2.2. Design the DHSA-AIFI module

As the core component of RT-DETR, AIFI module weights and fuses multi-scale features through adaptive attention mechanism to improve the accuracy and robustness of target detection. However, when dealing with complex backgrounds (such as lighting changes, complex textures) and multi-scale targets for cloth defects, multi-head self-attention in the AIFI module faces challenges: differences in visual characteristics caused by lighting and weather changes make multi-head self-attention difficult to adapt, especially when the target and background colors are similar, which may lead to information loss. In addition, traditional self-attention relies too much on local features, which makes it difficult to deal with occluded objects and affects the integrity of objects. In multi-scale target processing, traditional mechanisms are difficult to dynamically adjust the weights of different scale targets. Small-scale targets are easily submerged by large-scale targets, and the information interaction between scales is insufficient, which reduces the detection accuracy. Based on this, this study introduces Dynamic-range Histogram Self-Attention (DHSA) into AIFI module, and designs DHSA-AIFI module, whose structure diagram is shown in Figure 6. DHSA-AIFI can effectively capture long-distance semantic correlation in cross-layer features, improve cross-scale feature fusion, suppress background noise interference, and enhance the processing ability of multi-scale targets and complex backgrounds.

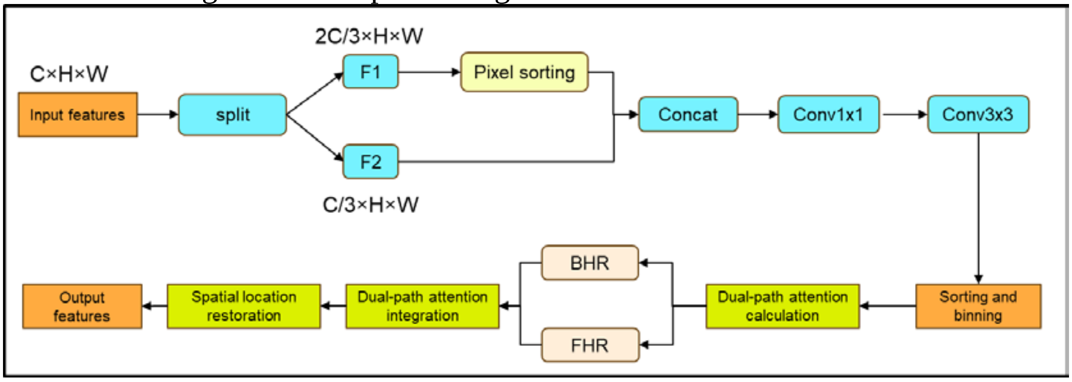


Figure 6: DHSA-AIFI.

The core of DHSA is to use binning logic to achieve long-distance similar feature aggregation, dual-path design to balance global context and local details, and introduce dynamic range convolution to enhance non-local feature interaction. First, the input data is subjected to feature extraction at different scales. The extracted feature map $F \in R^{C\times H\times W}$ is divided into two branches along the channel dimension, the main branch F1 and the auxiliary branch F2, occupying 2C/3 channels and C/3 channels respectively. F1 is sorted horizontally (Sorth) and vertically (Sortv) so that pixels of similar intensity are clustered spatially to obtain F_1^{sorted}.

$$F_1^{sorted} = Sort_v(Sort_h(F_1)) \tag{5}$$

The sorted F1 and F2 are spliced, the channel range is compressed by 1×1 point-by-point convolution, and then the features across the dynamic range are extracted by 3×3 depth convolution.

$$F_{drconv} = Conv_{3x3}^{d}(Conv_{1x1}(Concat(F_1^{sorted}, F_2))) \tag{6}$$

Where $Concat$ is the channel dimension splicing, $Conv_{1x1}$ compresses the number of channels to C. Second, the self-attention module is replaced by a two-path histogram self-attention, the output F_{drconv} is convolved over the dynamic range, sorted by pixel intensity and generated with index d, reshaped into a one-dimensional sequence:

$$V, d = Sort(R_{C\times H\times W}^{C\times HW}(F_{drconv})) \tag{7}$$

Where $R_{C\times H\times W}^{C\times HW}$ is the feature vectorization operation, $V \in R^{C\times HW}$, and then the one-dimensional sequence is divided into B bins sorted from high to low pixel intensity values, each bin contains $S = HW / B$ pixels, and the feature shape is $C\times B\times S$. Then attention fusion is performed by using two-path attention, which is divided into global interval attention (BHR) and local frequency attention (FHR).

In the global interval attention, Q1, K1 are arranged by sorting index d:

$$Q_1, K_1 = Split(Gather(F_{QK,1}, d)) \tag{8}$$

Q1 and K1 reshape to box dimension $C\times B\times S$, calculate inter-box attention:

$$A_B = soft\max(\frac{R_B(Q_1)\cdot R_B(K_1)^{\mathrm{T}}}{\sqrt{k}})\cdot R_B(V) \tag{9}$$

Where R_B is the box dimension remodeling and k is the number of attention heads.

In local frequency attention, Q2, K2 are also arranged by index d:

$$Q_2, K_2 = Split(Gather(F_{QK,2}, d)) \tag{10}$$

Q2 and K2 are remodeled to frequency dimension $C\times S\times B$, calculate attention in box:

$$A_F = soft\max(\frac{R_F(Q_2)\cdot R_F(K_2)^{\mathrm{T}}}{\sqrt{k}})\cdot R_F(V) \tag{11}$$

Where R_F is the frequency dimension remodeling, and each bin contains only B adjacent intensity pixels.

Finally, attention fusion is performed, and the resulting elements of the two paths are multiplied to restore the original spatial order:

$$A = Unsorted(A_B \odot A_F, d) \tag{12}$$

Where $Unsorted$ restores the feature to its original spatial location by index d.

2.3. Improve the cross-scale feature fusion module

RT-DETR uses CCFF feature pyramid structure for cross-scale feature fusion. First, the output feature channels are mapped to 256 through convolutional layer, and then multi-scale feature integration is realized through top-down and bottom-up bidirectional pairing fusion. However, CCFF is obviously inadequate in detecting small objects with cloth defects. Because of the small defect size, the traditional pyramid structure relies on high-level features (P3, P4, P5) for detection, but downsampling leads to the loss of detail information, which can not fully express the small target features, and is easy to miss or misdetect. In addition, defects are similar in color to background texture, high-level semantic information is difficult to distinguish between target and background, illumination changes and occlusion further weaken target edges and details, and reduce detection performance. Although adding P2 layer can supplement small target information, it will significantly increase the calculation amount and post-processing complexity, and it is difficult to meet the real-time detection requirements.

In order to solve these problems, SCOKCCFF feature pyramid structure is proposed in this study, which aims to optimize the feature expression of small targets, improve the discrimination ability between target and background, reduce the computational burden and improve the efficiency of multi-scale feature fusion. SCOK-CCFF is structured as shown in Figure 7 and consists of two core modules: spatial pyramid diluted convolution (SPDConv) and cross stage partial omnikernel (CSP-OmniKernel).

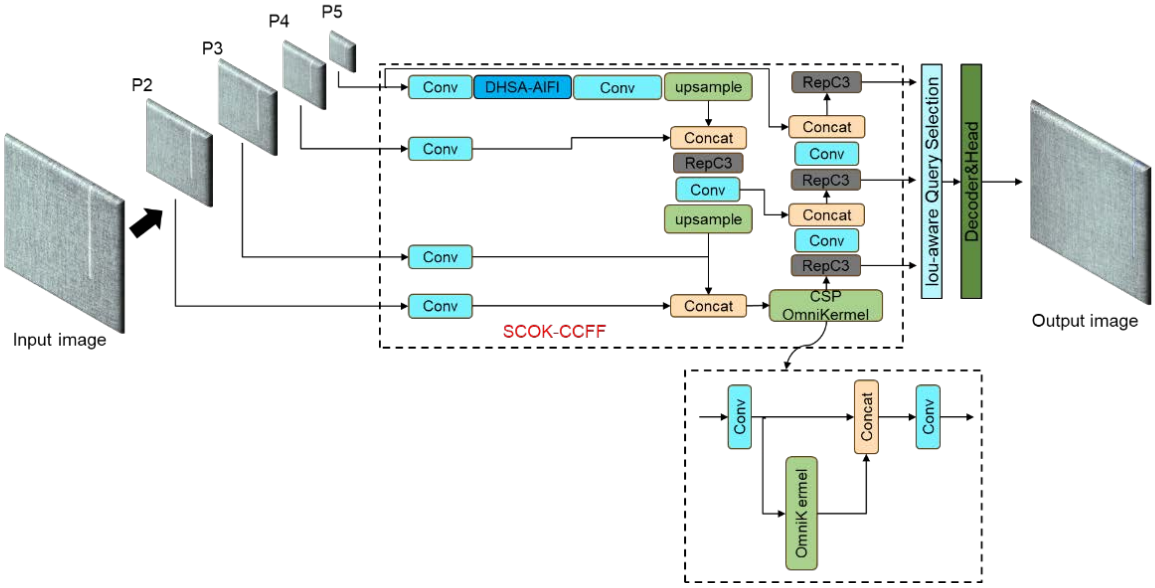


Figure 7: SCOK-CCFF network structure.

SPDConv module is introduced after P2 feature layer to solve the problem of insufficient detail expression of small targets. The P2 layer has high resolution and contains rich details of small targets, but direct use for detection will lead to a surge in computational effort. SPDConv efficiently processes P2 features through lightweight convolution operations, de-noising and extracting key small target information, which is then upsampled and passed to the P3 layer. In P3 layer, the optimized P2 features and P3 features are fused through Concat, which effectively makes up for the deficiency of P3 layer in small target expression, and avoids the computational burden brought by directly introducing P2. For the fusion of P3 and P2, traditional stitching or convolution operations cannot fully model the relationship between features at different scales, resulting in information redundancy and inefficient expression. Therefore, the CSP-OmniKernel module is introduced into the fused features for deep integration. CSP-OmniKernel module draws lessons from CSP (Cross Stage Partial) design idea and divides input characteristics into two parts: one part is directly transmitted, and the other part enters OmniKernel module for in-depth processing. OmniKernel optimizes features through multi-level design of global, large and local branches: global branch utilization Feature Spatial Attention Module (FSAM)(dense channel attention module (DCAM) captures long-range context information to optimize the distinction between target and background; large branches extract medium-scale structures through 32×32 convolution to maintain semantic information integrity; local branches focus on edge details and morphological features of small targets through 1×1 and 32×1 convolution to achieve multi-scale feature expression and interaction. The collaborative design of SPDConv and CSP-OmniKernel module effectively solves the problems of detail loss, target and background discrimination difficulty and computational burden in cloth defect detection, and significantly improves the accuracy and real-time performance of small target detection in complex scenes.

3. Experiments and results

3.1. Experimental environment

The computer operating system used in this experiment is Windows11 64-bit, the processor model is i7- 11800H, the graphics card model is NVIDA GeForce RTX3060, the main frequency is 2.3GHz, 16G machine with RAM, CUDA version is 11.8, the programming language is Python 3.8, and the deep learning framework Pytorch2.0.1.

The initial learning rate is set to 0.01, the batch size is set to 16, the rounds are 200 rounds, the input image size is 640× 640, the weight attenuation coefficient is 0.00005, the IoU threshold is set to 0.5, and the confidence threshold is set to 0.8.

3.2. Data sets and preprocessing

Table 1: Fabric numerical label and corresponding defect name.

Numerical Label	Defect Name	Numerical Label	Defect Name
1	Hole	8	Coarse Weft
2	Dirty Mark	9	Jacquard Skip
3	Three Threads	10	Starch Lump
4	Knot	11	Warping Knot
5	Coarse Warp	12	Star Skip
6	Loose Warp	13	Broken Spandex
7	Broken Warp	14	Thick-thin Place

The dataset used in this experiment is derived from the fabric defect dataset of the Alibaba Cloud Tianchi Competition. This paper applies 14 different defect types, and the entire dataset consists of 5,913 images. To meet the input requirements of the RT-DETR model, all images have been adjusted to 640×640 pixels. To further increase the quantity of the dataset, we have expanded the dataset and performed data augmentation operations on the original dataset. The dataset was expanded to 8,000 images mainly through flipping, splitting and adding noise. The expansion of the dataset can effectively enhance the generalization ability of the model, thereby reducing the possibility of overfitting. During the model training period, we set the ratio of the training set, test set to validation set at 7:2:1 to ensure the effective learning and validation of the model. In this experiment, each defect category is assigned a numerical label, and the corresponding defect names are summarized in the Table 1.

3.3. Evaluation index

In order to test the improved model better, R (Recall), P (Precision), mAP (Mean average precision), prediction time of a single picture, Params and Giga floating point operations per second (GFLOPs) are selected as evaluation indexes of the model for detecting fabric defects. The equations are as follows: (13)~(15).

Accuracy represents the proportion of all prediction results including correct prediction results, as shown in Equation (13):

$$P = \frac{TP}{TP + FP} \times 100\% \tag{13}$$

Recall represents the proportion of all targets correctly predicted, as shown in equation (14):

$$R = \frac{TP}{TP + FN} \times 100\% \tag{14}$$

Average precision represents the average of accuracy for all categories:

$$AP = \int_0^1 P(R)dR \tag{15}$$

Where TP is the number of defects accurately identified as positive cases; FP is the number of defects incorrectly predicted as positive cases;FN is the number of defects incorrectly predicted as negative cases.

3.4. Ablation experiments

Three improvements are made to the original RT-DETR-r18 model: A: BasicBlock is replaced by C2f_AdditiveBlock; B: AIFI is replaced by DHSA-AIFI; C: SCOK CCFF feature pyramid is designed for multi-scale feature fusion. Each improved module was separately integrated into the original model to verify its effectiveness. The corresponding results are shown in the table, where √ indicates that the corresponding enhancement module is used, × indicates that it is not used, and the best index is indicated in bold.

As can be seen from Table 2, after replacing BasicBlock with C2f_AdditiveBlock, the model performance is significantly improved, P, R and AP50 are improved by 1.4%, 2.1% and 1.6% respectively, the detection speed is improved by 4.4 frames/s, and the number of parameters and floating point calculations are reduced by 23.6% and 14.7% respectively. The results show that the improvement of backbone network can effectively enlarge the receptive field of feature extraction and enhance the perception ability of multi-scale target features. For the improvement of AIFI, DHSA-AIFI makes P increase by 4.1%, R increase by 0.6%, parameter quantity and calculation quantity decrease slightly, but it improves the ability to capture fine features and effectively suppresses the interference of complex background. The parameters and floating-point operations increase slightly after SCOK-CCFF is added, which may be due to the introduction of shallow high-resolution feature information, which increases the learning difficulty of the model. However, the maximum performance improvement was also achieved, with R and AP50 increasing by 3.7% and 4.8% respectively, verifying its effectiveness in detecting fine fabric defects. Compared with RT-DETR-r18, P, R and AP50 of the improved RT-FDTR model are increased by 6.9%, 2.5% and 3%, parameters and computation are reduced by 20.6% and 8.7% respectively, detection speed is increased to 59.7, and a good balance is achieved among speed, accuracy and computation. These results verify the effectiveness of the improved model.

Table 2: Ablation experiment result.

Experiment number	A	B	C	P/%	R/%	AP50/%	FPS	Params/M	Flops/G
1	×	×	×	75.3	74.6	73.5	50.0	19.9	57.3
2	√	×	×	76.7	76.7	75.1	54.4	**15.2**	48.9
3	×	√	×	79.4	75.2	74.8	53.1	20.1	58.4
4	×	×	√	81.6	**78.3**	**78.3**	56.5	21.3	58.1
5	√	√	√	**82.2**	77.1	76.5	**59.7**	15.8	52.3

3.5. Contrast experiment

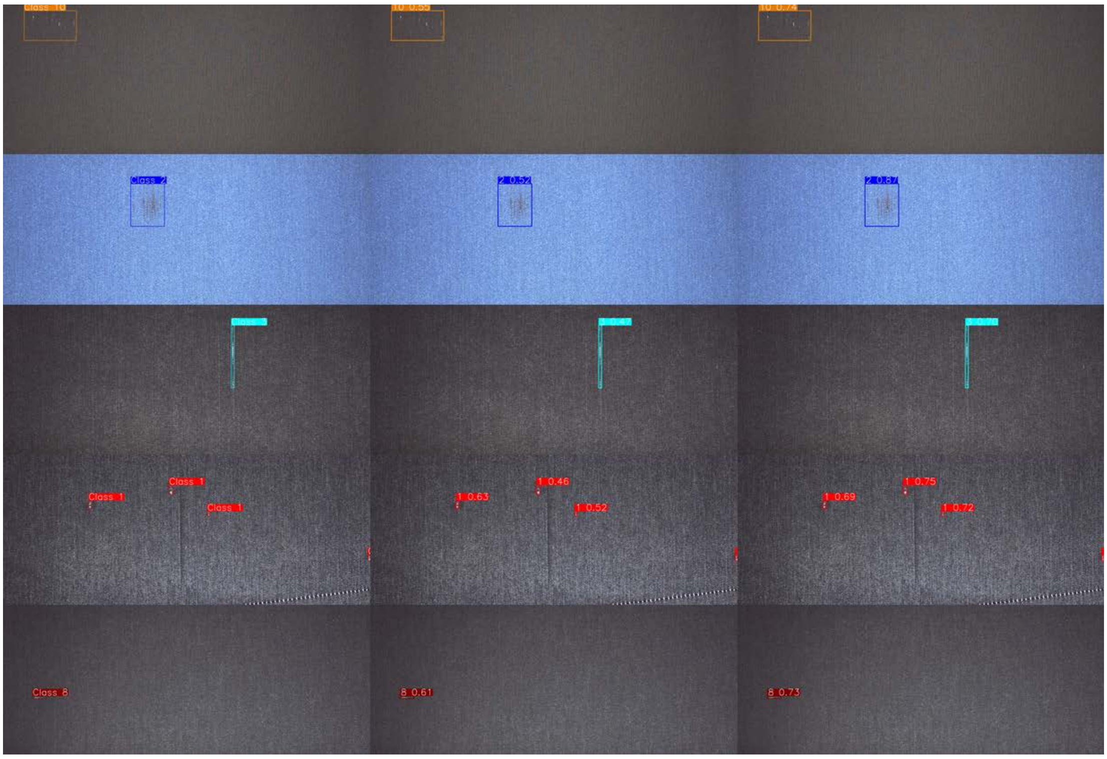
In order to further verify the performance of the improved model, this study selects seven advanced target detection algorithms to conduct comparative experiments on the same dataset. These seven algorithms include lightweight models (YOLOv5, YOLOv6, YOLOv8, YOLOv11) and network models based on Transformer architecture (RT-DETR-r18, RT-DETR-r34) as well as Faster R-CNN. The performance results for each network architecture are shown in Table 3.

Table 3: Comparative experiment result.

model	P/%	R/%	AP50/%
YOLOv5	64.2	59.3	63.6
YOLOv6	65.4	60.2	62.8
YOLOv8	74.5	72.2	72.1
YOLOv11	73.3	70.0	70.6
RT-DETR-r18	75.3	74.6	73.5
RT-DETR-r34	77.4	75.2	75.6
Faster R-CNN	60.3	58.7	62.9
RT-FDTR	82.2	77.1	76.5

From the data in Table 2, it can be seen that the recognition accuracy and recall rate of the improved model based on RT-DETR proposed in this paper reach 82.2% and 77.1% on the dataset, and $mAP_{0.5}$ reaches 76.5%. For AP50, compared with YOLOv5, YOLOv6, YOLOv8, YOLOv11, Fastet R-CNN, the recognition accuracy and recall rate are improved by 12.9%, 13.7%, 4.4%, 5.9% and 13.6% respectively. This is because the features extracted by CNN have limited modeling ability for global context information, while the improved algorithm can extract more detailed shallow features. Compared with RT-DETR-r18 and RT-DETR-r34, the AP50 of the improved model increased by 3% and 0.9% respectively. Compared with other mainstream networks, the improved model based on RT-DETR has better algorithm performance, which shows that the improved model can extract feature information better in the case of complex interference, locate the target region, and thus improve the recognition accuracy and recall. The comprehensive performance is better than other mainstream target detection models.

As shown in Figure 8, the detection results of the original RT-DETR-r18 model and RT-FDTR model are represented by rectangular boxes, and the category labels and confidence levels of recognition are marked.



(a) (b) (c)

Figure 8: Detection result comparison: (a) ground truth. (b) RT-DETR detect result. (c) RT-FDTR detect result.

Figure 8 (b) shows that the RT-DETR-r18 model has poor detection ability for cloth defects; Figure 8 (c) shows that the RT-FDTR model has greatly improved detection effect for cloth defects compared with the original model, further proving the effectiveness of the improved model.

4. Conclusions

In this paper, a lightweight fabric defect detection network RT-FDTR is proposed to solve the problems of traditional fabric defect plus memory algorithm, such as low accuracy, poor real-time performance and weak adaptability to complex scenes. Firstly, AdditiveBlock and CSP in CAS-ViT are introduced to optimize the backbone of the network, expand the receptive field, and improve the ability of extracting global texture and multi-scale defect features of cloth. Secondly, a dynamic range histogram self-attention mechanism is designed to realize cross-layer long-distance semantic association capture through dual-path structure, which strengthens the ability to capture details of small targets and complex backgrounds, suppresses noise interference, and improves the adaptability of the model to multi-scale targets.

Finally, SCOK-CCFF feature pyramid is developed, and small target information is extracted through SPD-Conv denoising. Combined with CSP-OmniKernel deep integration feature, SCOK-CCFF feature pyramid is proposed, which significantly improves small target detection accuracy and background discrimination ability. The researchers use the improved RT-FDTR to analyze cloth data sets.

Finally, the recognition rate, recall rate and average accuracy are 82.2%, 77.1% and 76.5% respectively, which are improved by 6.9%, 2.5% and 3% compared with the original model, and the parameter quantity is reduced by 20.6%, and the calculation speed is improved by 9.7 frames/s, which meets the requirements of industrial production for cloth defect detection. The RT-FDTR fabric defect detection method proposed in this study can meet the actual production requirements.

References

[1] Jia X J, Ye L H, Deng H T, et al. Classification of primitive patterns of blue calico based on convolutional neural network[J]. Journal of Textile Research, 2020, 41(1): 110-117.

[2] WANG X B, FANG W J, XIANG S. Fabric defect detection based on anchor - free network [J]. Measurement science and technology, 2023, 34(12): 12.

[3] Zhang L, Zhu W J, Zhu S W. Progress in automatic detection methods and applications of fabric defects[J]. Progress in Textile Science & Technology, 2022(2): 21-26.

[4] Zhang K X, Du J L. Fabric defect detection method based on improved YOLOv5[J]. Modern Electronics Technique, 2024, 47(20): 109-117.

[5] Liu Zhoufeng, Tian Bo, Li Chunlei, Ding Shumin & Xi Jiangtao. CACFNet: Fabric defect detection via context-aware attention cascaded feedback network.Textile Research Journal, 2023, 93(13-14), 3036-3055.

[6] Chen C ,Zhou Q ,Li S , et al.Fabric defect detection algorithm based on improved YOLOv8[J].Textile Research Journal,2025,95(3-4):235-251.

[7] M. Zhang, W. Yu, H. Qiu, J. Yin and J. He, "A Fabric Defect Detection Algorithm Based on YOLOv8," 2023 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML), Chengdu, China, 2023, pp. 1040-1043.

[8] S. S. Mohammed and H. G. Clarke, "Advanced Convolutional Neural Network Approach for Fabric Defect Detection,"2024 Innovations in Intelligent Systems and Applications Conference (ASYU), Ankara, Turkiye, 2024, pp. 1-5.

[9] L. Yao, S. Song and Y. Wan, "Fabric Defect Detection Based on Hybrid Attention Transformer and Improved Cascade R-CNN," 2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Kuching, Malaysia, 2024, pp. 1933-1938.

[10] Zhang Hongwei, Qiao Guanhua, Lu Shuai, Yao Le & Chen Xia. Attention-based Feature Fusion Generative Adversarial Network for yarn-dyed fabric defect detection.Textile Research Journal, 2023, 93(5-6), 1178-1195.

[11] Qin J H, Chen Z L, Wan B X, et al. Green orange detection method in complex orchard environment based on RT-DETR[J]. Electronic Measurement Technology, 2025, 48(11): 175-186.

[12] Y. Zhao et al., "DETRs Beat YOLOs on Real-time Object Detection,"2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2024, pp. 16965-16974.