Available Online: 30 April 2025 DOI: 10.54254/2977-3903/2025.22684

Target localization of abrasion-resistant color fastness samples based on YOLOv8 optimization and enhancement

Yaling Cao

College of Electrical and Electronic Engineering, Wenzhou University, Wenzhou, China

cyl1911568117@163.com

Abstract. To address the challenges in detecting abrasion-resistant color fastness samples – including limited sample instances, non-uniform shapes, and insufficiently distinct texture variations that compromise localization accuracy – this paper optimizes the detection framework through the integration of three key strategies: Global Attention Mechanism (GAM), Dynamic Sampling (DySample), and Adaptively Spatial Feature Fusion (ASFF), thereby enhancing detection accuracy and efficiency. Initially, Mosaic data augmentation is implemented to enrich dataset diversity and improve model robustness. Subsequently, the GAM attention mechanism is embedded into the backbone network to enhance target feature extraction capabilities. DySample replaces conventional upsampling methods in the neck network to achieve more effective feature reconstruction. Finally, the ASFF module is integrated into the Detect module within the head network to enable adaptively spatial weight learning for multi-scale feature map fusion. Compared with baseline algorithms, the improved framework demonstrates performance gains of 1.2% in Precision, 3.0% in Recall, 1.2% in mAP@0.5, and 13.5% in mAP@0.5:0.95. Experimental results validate the effectiveness of the proposed method, which maintains satisfactory performance across additional datasets, demonstrating strong robustness and superior generalization capability.

Keywords: target detection, YOLOv8, abrasion-resistant color fastness sample, convolutional neural network

1. Introduction

The quality of textile products significantly impacts consumer experience and personal health. As one of the critical indicators for evaluating textile dyeing quality, color fastness reflects the color stability of textiles during use or washing processes. Rapid and accurate assessment of color fastness levels holds substantial significance for the textile printing and dyeing industry [1]. Compared with traditional visual evaluation methods, modern instrumental assessment demonstrates greater scientific objectivity. The mainstream evaluation instrument—digital colorimeters—has evolved significantly. The non-contact color measurement instrument DigiTool [2] offers enhanced operational convenience compared to other devices. By capturing sample images using its built-in digital camera and anchoring targets, the integrated Digi-Grade evaluation software automatically grades color fastness samples. Abrasion-resistant samples are more difficult to detect compared to multi-fiber samples such as soap-fastness color fastness, and there are fewer sample instances than multi-fiber samples.

To address the subjectivity and labor-intensive nature of manual color fastness evaluation, An [3] proposed an approach combining Canny operator-based edge detection for feature region identification with minimum bounding rectangle extraction and segmentation. This method integrates color measurement instruments with deep learning to reduce operational requirements while maintaining simplicity and efficiency, though detection accuracy remains improvable. Liu [4] enhanced small target detection accuracy while reducing parameters by incorporating BiFormer attention mechanisms and GSConv modules into YOLOv8. Jiang [5] developed AEM-YOLOv8s, introducing a C2f-BE module that combines AKConv with EMA attention mechanism to improve small target detection. For complex cotton field environments, Zheng [6] proposed the YOLOv8-DMAS model, replacing all BottleNeck layers with dilated residual modules in C2f networks, integrating multi-scale modules (MSBlock) in the final backbone layer, and enhancing detection heads with Adaptively Spatial Feature Fusion (ASFF) alongside SoftNMS replacement. Wu [7] advanced the YOLOv8n algorithm by integrating SPD-Conv modules, GAM attention mechanisms, and Wise-IoU loss functions to optimize apple detection and segmentation in occlusion scenarios (SGW-YOLOv8n). Despite rapid advancements in object detection algorithms that have significantly enhanced model prediction precision, their application in color fastness sample localization remains underexplored.

Given YOLOv8's rapid iteration, widespread application, stable architecture, high precision, and compact parameterization, this study selects YOLOv8n as the baseline network for abrasion-resistant sample localization. Through integration of Mosaic augmentation, global attention mechanism, dynamic sampler, and adaptively spatial feature fusion, the optimized framework achieves enhanced detection accuracy and operational efficiency.

2. Enhancement strategy

In this research, YOLOv8 [8] serves as the baseline model, comprising core components: input images (Input), feature extraction backbone (Backbone), feature fusion network (Neck), and prediction module (Head). The overall network architecture is illustrated in Figure 1. In this study, YOLOv8 was used as the baseline model, comprising essential components: input images, a feature extraction backbone, a feature fusion network (the neck), and a prediction module (the head). The enhancement strategies in this paper focus on improvements in the GAM attention mechanism within the backbone, DySample in the neck, and ASFF in the detection head, addressing issues such as the scarcity of instances for color fastness under friction, difficulty in target localization, and the challenge of meeting high detection accuracy requirements. By integrating the GAM attention mechanism into the SPPF module of the Backbone, feature information can be extracted more effectively; replacing UpSample in the Neck with the dynamic upsampling method DySample results in lower inference latency and lighter weight; and the ASFF fusion in the Head module facilitates better integration of feature layers at different scales, thereby enhancing detection accuracy.

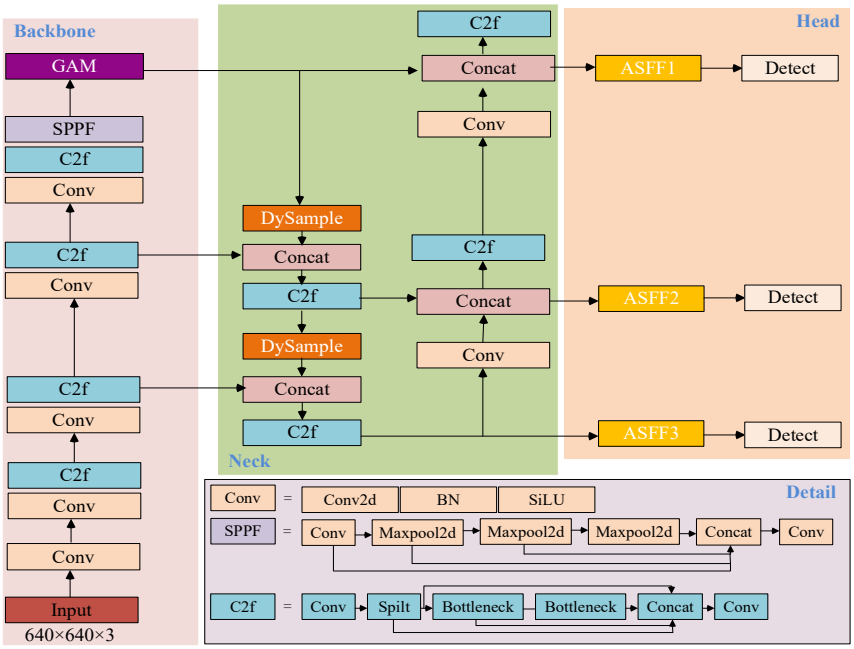

Figure 1. The overall structure of the network

2.1. Retain information to enhance channel-spatial interactions (GAM)

To better focus on critical information in images, researchers have maintained strong interest in exploring attention mechanisms. Currently, the most widely used attention mechanisms are the Squeeze-and-Excitation attention mechanism (SE) [9] and the Convolutional Block Attention Module attention mechanism (CBAM) [10]. However, SE demonstrates limited efficiency in suppressing unimportant pixels, while CBAM neglects channel-spatial interactions. The Global Attention Mechanism (GAM) [11] mitigates information loss caused by pooling in CBAM and amplifies global-dimensional interactive features, thereby addressing the insufficient retention of information across both channel and spatial dimensions in traditional attention mechanisms. Figure 2 illustrates the schematic of the GAM module. For a given input feature map $F_1 \in R^{C\times H\times W}$, the intermediate states F_2 and the output F_3 are defined as:

$$F_2 = M_C(F_1) \otimes F_1 \tag{1}$$

$$F_3 = M_S(F_2) \otimes F_2 \tag{2}$$

where M_C, M_S are channel and spatial attention diagrams, $\otimes$ representing the multiplication of elements.

The global attention mechanism comprises two core sub-modules: the Channel Attention Module (CAM) and the Spatial Attention Module (SAM). The CAM employs a 3D permutation architecture to preserve three-dimensional information integrity.

Given an input feature map $F_1 \in R^{C\times H\times W}$, the CAM first reorganizes tensor dimensions to generate $F_1 \in R^{W\times H\times C}$, effectively redistributing channel-spatial relationships. Subsequently, a two-layer Multilayer Perceptron (MLP) with compression ratio r performs dimensionality reduction followed by reconstruction, enhancing cross-dimensional channel-space correlations. The refined features are then processed through a sigmoid activation function to produce channel-wise attention weights. The SAM further optimizes spatial information extraction by integrating two 7×7 grouped convolutional layers, which systematically aggregate contextual patterns while maintaining computational efficiency. Given that maximum pooling operations can result in a loss of information that negatively impacts performance, the module discards pooling operations to ensure that the integrity of the feature mapping is preserved.

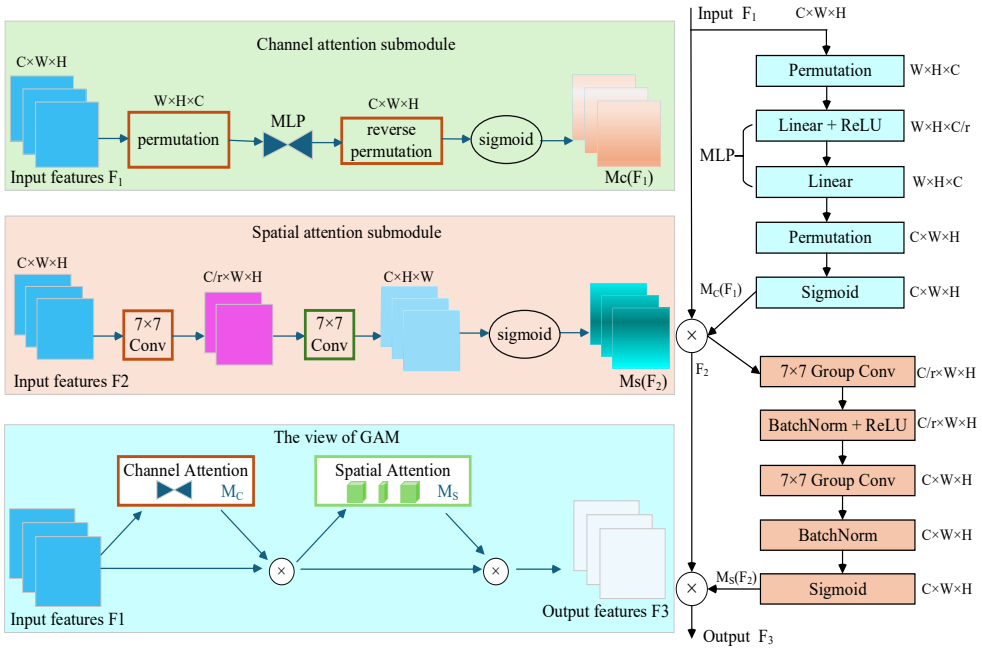

Figure 2. The global attention mechanism integrates channel and spatial attention module

2.2. Adaptively Spatial Feature Fusion (ASFF)

The traditional YOLOv8 model utilizes a feature pyramid network structure, achieving the adaptive weighted fusion of multi-scale features through a bidirectional feature fusion method that is both top-down and bottom-up. However, this multi-scale fusion mechanism has significant limitations: conflicts of information can arise between features at different levels, leading to inconsistencies in feature expression and severely restricting improvements in detection performance. To address this issue, this paper introduces the adaptively spatial feature fusion [12] module. This module effectively suppresses information conflicts during the cross-scale feature fusion process through a spatial dimension feature filtering mechanism. The core advantage of the ASFF module lies in its ability to enhance the scale invariance of features, thereby significantly improving the performance of the feature pyramid, e.g. Figure 3.

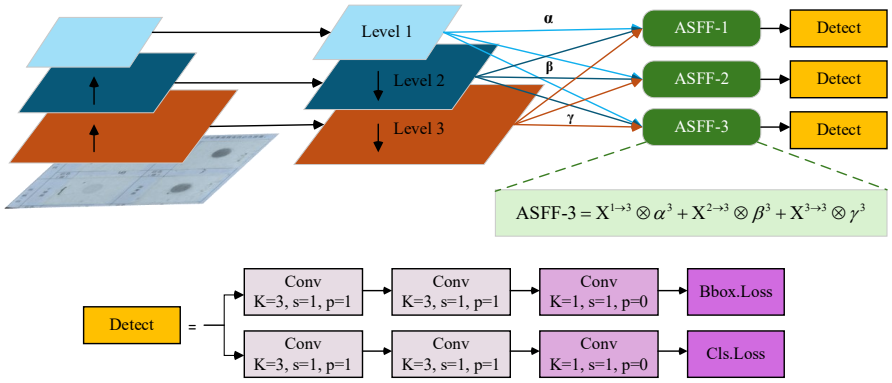

Figure 3. The adaptively spatial feature fusion is introduced into detect in the head

Level 1, Level 2, and Level 3 refer to the three feature maps output from the Neck layer of the YOLOv8 model. The key idea of ASFF is to adaptively learn the fusion spatial weights of each scale feature map, which is divided into two steps: identity scaling and adaptively fusion. Identity scaling includes 1/2 downsampling, 1/4 downsampling, and upsampling operations. Let the features of level l be denoted as y^l, for level l, the feature extraction maps y^n of other levels n $(n \neq l)$ are scaled to the same size as y^l. Taking ASFF-3 as an example, $X^{2\to3}$ can be obtained by performing a 1×1 convolution operation on the feature map of level-1 and resize it to 4 times the resolution of the original image, and $X^{1\to3}$ can be obtained by performing a 1×1 convolution operation on the feature map of level-2 and resize it to 2 times the resolution of the original image. The adaptive fusion combines feature outputs from different layers multiplied by the learnable weighting coefficients, α^3, β^3 and γ^3, and then sums them to obtain new fused features. Taking ASFF-3 as an example, see the following formula:

$$y_{ij}^{l} = \alpha_{ij}^{l} \cdot X_{ij}^{1\to l} + \beta_{ij}^{l} \cdot X_{ij}^{2\to l} + \gamma_{ij}^{l} \cdot X_{ij}^{3\to l} \tag{3}$$

where is the output feature map (i, j) vectors between channels; α_{ij}^{l}, β_{ij}^{l}, γ_{ij}^{l}, are the learnable weights of the feature graphs of three different levels, and satisfy $\alpha_{ij}^{l} + \beta_{ij}^{l} + \gamma_{ij}^{l} = 1_{j}$, $\alpha_{ij}^{l}, \beta_{ij}^{l}, \gamma_{ij}^{l}, \in [0,1]$; $X_{ij}^{1\to l}$, $X_{ij}^{2\to l}$, $X_{ij}^{3\to l}$ is the output result of a location feature map.

2.3. Lightweight dynamic upsampling module (DySample)

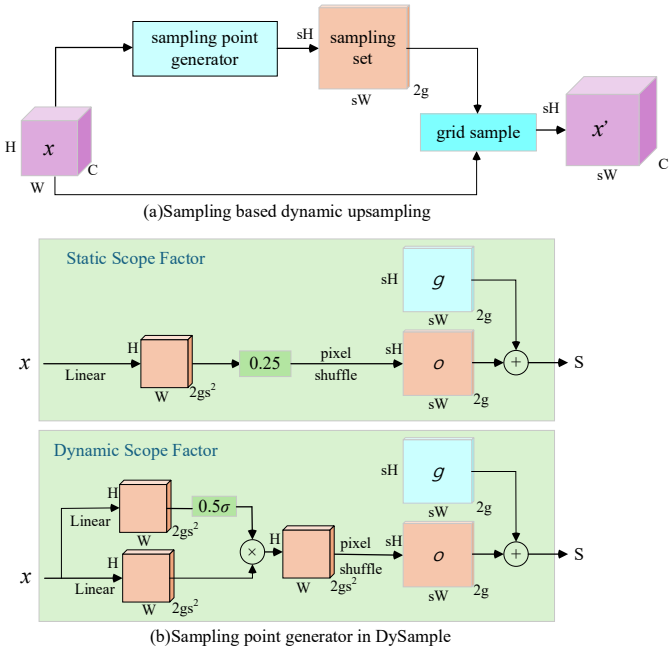

Figure 4. Dynamic upsampling and static, dynamic point samplers

Upsampling is a key component of the YOLOv8 network architecture, as it allows low-resolution features to be elevated to the same scale as high-resolution features, enabling the model to precisely localize and identify object details. This study introduces the dynamic upsampling [13] module in the neck network, which is a lightweight and efficient dynamic upsampling operator. By retaining only significantly different pixel values, it can effectively reduce data volume, lower computational complexity, and enhance upsampling efficiency. From a point sampling perspective, DySample dynamically adjusts sampling offsets based on input feature map content, not only improving model robustness against interference but also conserving computational resources, making it suitable for industrial real-time detection. Figure 4 illustrates the feature map processing process using the dynamic factor sampling method in DySample. As shown in Figure 4(a), a feature map χ of a specified size $C \times H_1 \times W_1$ and a point sampling set $2g \times sH \times sW$ as input, where the first dimension of $2g$ represents the x and y coordinates, and the feature map χ is resampled by the grid_sample function to χ' of size $C \times sH \times sW$ using the position in the point sample set S.

$$\chi' = \text{grid\_sample}(\chi, \delta) \tag{4}$$

The process by which the sample point generator generates sample set P is shown in Figure 4(b), which provides two ways to optimize the sampling quality: one is to introduce a static range factor (typically 0.25) to limit the range of the offset and reduce overlap. The second is the introduction of a dynamic range factor, which is obtained by multiplying the standard linear operation

with another linear operation using 0.5σ, allowing the model to dynamically adjust the offset range at each point based on the content of the feature map. First, the feature map χ generates an offset O with the size $2gs^2 \times H \times W$ through a linear operation:

$$O = linear(\chi) \tag{5}$$

Then reshape the high-resolution original sample grid G with size $2g \times sH \times sW$ by pixel shuffle, then the point sample set S is the sum of the offset O and the original sample grid G:

$$S = G + O \tag{6}$$

3. Experimental investigation and results analysis

3.1. Experimental dataset and evaluation metrics

The dataset utilized in this study was obtained from abrasion-resistant color fastness samples provided by SCOCIE Company (Zhejiang, China). As illustrated in Figure 5, these samples were annotated using LabelImg software, encompassing four friction categories: longitudinal dry friction (friction_point0), longitudinal wet friction (friction_point1), transverse dry friction (friction_point2), and transverse wet friction (friction_point3). To expand the dataset, the original images underwent data augmentation techniques including random rotation, horizontal vertical flipping, and scaling, resulting in 5,460 annotated images, which were partitioned into training, validation, and test sets in a ratio of 8:1:1. Select Adaptive Moment Estimation with Weight Decay (AdamW) as the optimizer for training, with an initial learning rate of 0.00125, a batch size of 16, and an input resolution of 640×640 pixels over 300 epochs. Other experimental environment setup are presented in Table 1. Network performance was evaluated using metrics including Precision, Recall, Average Precision (AP), mean Average Precision (mAP), and parameter counts. Specifically, mAP@0.5 denotes the mean average precision at an Intersection-over-Union (IoU) threshold of 0.5, while mAP@0.5:0.95 represents the average precision computed across IoU thresholds ranging from 0.5 to 0.95 in 0.05 increments.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{7}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{8}$$

$$\text{AP} = \int_0^1 p(r)dr \tag{9}$$

$$\text{mAP} = \frac{1}{n}\sum_{k=1}^{k=n} AP_k \tag{10}$$

where TP is the number of positive samples judged to be positive, FP is the number of negative samples judged to be positive, FN is the number of positive samples judged to be negative, n is the number of classes, and $p(r)$ is the Precision-Recall curve.

Table 1. Experimental environment setup

Experimental platform	Experimental configuration
Operating system	Ubuntu18.04
GPU	GEFORCE RTX 4090D(24GB)
CPU	Intel(R) Xeon(R) Platinum 8474C
Deep learning network acceleration library	Cuda 11.1
Programming language	Python3.8
Network framework	Pytorch 1.9.0

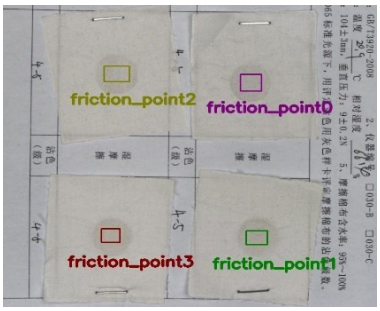

Figure 5. Annotation of abrasion-resistant samples

3.2. Mosaic data augmentation

The Mosaic Data Augmentation [14] method is proposed in the YOLOv4 paper, referring to the CutMix data augmentation algorithm, the main idea is to randomly crop four pictures and then stitch them onto a graph as training data, so that the model can identify targets in a smaller range, which greatly enriches the background of the detected object. Firstly, the connection base point of the image is randomly selected, and then 4 images are randomly selected, and the size is adjusted and scaled respectively according to the coordinates of the reference point, and it is placed in the upper left, upper right, lower left and lower right of the large picture of the specified size; Then, according to the transformation mode of each image, the mapping relationship corresponding to the image annotation is displayed in the adjusted picture. Finally, according to the determined horizontal and vertical coordinates, the adjusted images are spliced and the coordinates of the detection frame exceeding the boundary are processed.

As can be seen from Table 2, this enhancement resulted in 0.8% improvement in Persion, 0.7% improvement in Recall, 0.5% decrease in map@0.5, and 7.7% improvement in map@0.5:0.95. The results highlight that the Mosaic data augmentation can effectively improve the object detection performance, especially the average accuracy at higher thresholds is significantly improved, indicating that the model's detection ability for multi-scale targets is greatly improved, especially in small targets and occlusion scenarios.

Table 2. Mosaic data augmentation effect comparison experiment

Experimental Method	Parameters /M	Precision (%)	Recall (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)
YOLOv8	3.01	95.9	93.2	97.2	56.9
YOLOv8-Mosaic	3.01	96.7	93.9	97.5	64.6

3.3. Comparative experiment with attention modules

In order to fully validate the effectiveness of the GAM module on the YOLOv8 target detection algorithm, it was decided to incorporate the CBAM, EMA, and SE attention modules into the network for comparative experiments. The experimental results are shown in Table 3, which indicates that SE and EMA are not suitable for this dataset, as their inclusion in the network resulted in contrary effects. Both CBAM and GAM demonstrate varying improvements in the experiments; the former exhibits increases in both accuracy and recall, while the latter shows a slight decrease in accuracy but a significant increase in recall, effectively reducing the missed detection rate, which holds significant importance in the field of industrial inspection. Furthermore, the average precision is improved more than that of the attention mechanism of CBAM, especially at high confidence thresholds.

Table 3. Effect of different attention mechanisms on outcomes

Experimental Method	Parameters /M	Precision (%)	Recall (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)
YOLOv8-Mosaic	3.01	96.7	93.9	97.5	64.6
YOLOv8-Mosaic-SE	**3.01**	96.2	93.7	97.1	63.8
YOLOv8-Mosaic-EMA	3.02	96.4	94.2	97.5	63.7
YOLOv8-Mosaic-CBAM	3.02	**96.8**	94.3	97.6	64.2
YOLOv8-Mosaic-GAM	3.48	96.0	**94.7**	**97.7**	**65.2**

3.4. Ablation experiments

In order to validate the feasibility of each module in the improved strategy, ablation experiments were conducted on each enhancement module using the baseline model to verify the superiority of the methods presented in this chapter. DySample replaced the upsampling in the Neck, the ASFF enhancement module was added to the Detect section of the Head, and the GAM module was incorporated into the SPPF module of the Backbone. The effectiveness of different modules on dataset enhancement was analyzed using differential ablation experiments, the results of which are presented in Table 4. After employing Mosaic data augmentation, mAP@0.5:0.95 improved by 7.7%, significantly enhancing detection performance and greatly assisting the model's generalization capabilities. The inclusion of DySample improved recall rates and reduced false negatives, although it slightly decreased localization accuracy at high thresholds. Following the addition of the ASFF module, there were significant improvements in Recall, mAP@0.5, and mAP@0.5:0.95, indicating that the ASFF module effectively integrates multi-scale features and enhances detection performance. The incorporation of the GAM attention mechanism further optimized detection

performance. Compared to the improved network, Precision, Recall, mAP@0.5, and mAP@0.5:0.95 increased by 1.2%, 3.0%, 1.2% and 13.5%.

Table 4. Impact of different modules on network performance

YOLOv8	Mosaic	DySample	ASFF	GAM	Precision (%)	Recall (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)
√					95.9	93.2	97.2	56.9
√	√				96.7	93.9	97.5	64.6
√	√	√			96.1	94.5	97.8	62.9
√	√	√	√		97.0	95.5	98.2	69.2
√	√	√	√	√	**97.1**	**96.2**	**98.5**	**70.4**

3.5. Comparison experiments among mainstream algorithms

In the comparative experiments, the method we proposed was evaluated against established object detection algorithms, including Single Shot MultiBox Detector (SSD) [15], Faster RCNN [16], YOLOv3, YOLOv5, YOLOv8, and YOLOv10, as shown in Table 5. Compared to other algorithms, the improved algorithm achieved better detection results, with the Average Precision (AP) of the four targets leading. There was an improvement of 6.9%, 3.9%, 0.8%, 1.4%, 1.3%, and 1.4% compared to SSD, Faster RCNN, YOLOv3, YOLOv5n, YOLOv8n, and YOLOv10n, respectively. In conclusion, the proposed enhancement strategy can effectively improve the detection accuracy of samples with resistance to abrasion-resistant color fastness when compared to object detection algorithms.

Table 5. Performance comparison of different methods

Experimental Method	AP/%				mAP@0.5 (%)	Parameters /M
	Friction_point0	Friction_point0	Friction_point2	Friction_point3		
SSD	86.2	95.5	89.2	95.5	91.6	24.6
Faster RCNN	90.9	97.6	93.2	96.8	94.6	36.7
YOLOV3	96.0	**99.6**	96.2	**99.2**	97.8	103.7
YOLOV5-n	93.9	98.9	95.6	98.7	96.8	**2.51**
YOLOV8-n	95.6	98.9	95.6	98.9	97.2	3.01
YOLOV10-n	95.0	98.9	96.3	98.1	97.1	2.71
Improved algorithm	**96.7**	99.4	**98.5**	99.0	**98.5**	4.86

3.6. Application of improved algorithms on other datasets

The VisDrone2019 dataset [17] is a large-scale dataset from a drone's perspective, open-sourced by teams including Tianjin University. To verify whether the improved algorithm performs similarly on other datasets, it was chosen to apply the improved algorithm to the VisDrone2019 dataset. The experimental results are shown in a Table 6. Precision, Recall, mAP@0.5, and mAP@0.5:0.95 have respectively increased 1.9%, 1.5%, 2.1%, 1.1%. The improved network model also demonstrates excellent performance on other datasets, proving that the improved algorithm has better robustness and generalization.

Table 6. Application of improved algorithms in VirDrone2019

Experimental Method	Precision (%)	Recall (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)
YOLOv8	42.5	31.3	30.7	17.8
Improved algorithm	44.4	32.8	32.8	18.9

4. Conclusion

In this paper, the object detection algorithm of rubbing fastness samples based on YOLOv8n network structure is optimized by integrating strategies such as global attention mechanism, dynamic sampling and adaptively spatial feature fusion, which

effectively improves the detection performance and reduces the labor cost. Using the dataset provided by Zhejiang SCOCIE, several experiments were carried out. Mosaic data augmentation significantly enhances the model's ability to detect multi-scale targets, especially improving the average accuracy at high thresholds, showing obvious advantages at high confidence thresholds. Ablation experiments verified the effectiveness of each improved module, DySample improved the recall rate and reduced the missed detection rate, ASFF effectively fused multi-scale features, greatly improved the performance of the detection target, and compared with other attention modules, the global attention mechanism GAM could optimize the detection results more effectively and further optimize the detection performance. Compared with the mainstream algorithms, the improved algorithm shows significant advantages in detection accuracy and is better than other algorithms. In addition, the improved algorithm was applied to the VisDrone2019 dataset to achieve the enhancement of various indicators, which proved its robustness and generalization ability.

In summary, the enhancement strategy proposed in this paper significantly improves the detection accuracy of the abrasion-resistant sample, and also shows good adaptability on other datasets, which has important application value and research significance. In the future, we will explore the application of algorithms in more complex scenarios and different types of datasets to continuously optimize the performance of algorithms.

References

[1] Zhang, H., Tang, Y., & Zhou, W. (2022). Introductory Analysis of Colour Fastness to Textiles. *Textile Testing and Standards, 8*(04), 21-24-32.

[2] Ruan, X., Lin, F., & Cui, G. (2022). Textile color fastness grading device (CN217277926U) [Utility Model Patent]. China National Intellectual Property Administration.

[3] An, Y., Xue, W., & Ding, Y., (2022). Grading of color fastness to rubbing of textiles based on image processing. *Journal of Textile Research, 43*(12), 131-137.

[4] Liu, Z. Y., Xu, H. Y., Zhu, X. Z., Li, C., Wang, Z. Y., Cao, Y. Q., & Dai, K. J. (2024). Bi-YOLO: An Improved Lightweight Object Detection Algorithm Based on YOLOv8n. *Computer Engineering and Science, 46*(08), 1444-1454.

[5] Jiang, W., Wang, W., & Yang, J. (2024). AEM-YOLOv8s: Small target detection in UAV aerial images. *Computer Engineering and Applications*, 60(17), 191-202.

[6] Zheng, L., Yi, J., He, P., Tie, J., Zhang, Y., Wu, W., & Long, L. (2024). Improvement of the YOLOv8 Model in the Optimization of the Weed Recognition Algorithm in Cotton Field. *Plants, 13*(13), 1843-1843.

[7] Wu, T., Miao, Z., Huang, W., Han, W., Guo, Z., & Li, T. (2024). SGW-YOLOv8n: An Improved YOLOv8n-Based Model for Apple Detection and Segmentation in Complex Orchard Environments. *Agriculture, 14*(11), 1958-1958.

[8] Ultralytics. (2023). YOLOv8 [Source code]. GitHub. Retrieved April 10, 2023, from https://github.com/ultralytics/ultralytics

[9] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7132-7141). IEEE. https://doi.org/10.1109/CVPR.2018.00745

[10] Woo, S., Park, J., Lee, J.Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), *Computer vision - ECCV 2018: 15th European Conference on Computer Vision, Munich, Germany, September 8-14, 2018, Proceedings, Part VII* (pp. 3-19). Springer. https://doi.org/10.1007/978-3-030-01234-2_1

[11] Liu, S., Huang, D., & Wang, Y. (2019). Learning Spatial Fusion for Single-Shot Object Detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. arXiv:1911.09516.

[12] Liu, W., Lu, H., Fu, H., & Cao, Z. (2023). Learning to upsample by learning to sample. *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 6004-6014). IEEE. https://doi.org/10.1109/ICCV51070.2023.00554

[13] Bochkovskiy, A., Wang, C., & Liao, H. M. (2020). YOLOv4: Optimal Speed and Accuracy of Object Detection. In *Computer Science - Computer Vision and Pattern Recognition*. arXiv:2004.10934.

[14] 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 (ECCV)* (pp. 21-37).

[15] Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(6), 1137-1149.

[16] Du, D., Zhu, P., Wen, L., Bian, X., Lin, H., & Hu, Q. (2019). VisDrone-DET2019: The Vision Meets Drone Object Detection in Image Challenge Results. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)* (pp. 213-226).

[17] Li, Y., Shi, Z., & Hoffmann (2021). Global Attention Mechanism: Retain Information to Enhance Channel-Spatial Interactions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. arXiv:2112.05561.