Electronics and Electrical Engineering

AIMS Electronics and Electrical Engineering, 10(1): 71–91.
DOI: 10.3934/electreng.2026004
Received: 18 July 2025
Revised: 15 September 2025
Accepted: 27 October 2025
Published: 14 January 2026

https://www.aimspress.com/journal/ElectrEng

Research article

CD-YOLO: A lightweight end-to-end detection model for cigarette appearance defects

Yuanyuan Liu, Hao Wu, Hao Zhou and Guowu Yuan*

School of Information Science and Engineering, Yunnan University, Kunming 650504, China

*** Correspondence:** Email: gwyuan@ynu.edu.cn; Tel: +86-871-65033748.

Abstract: Appearance defect detection is essential for ensuring cigarette quality during production. Reaching high-precision and lightweight automated cigarette appearance defect detection has long been manufacturers' key focus. However, existing methods struggle to balance detection accuracy and speed effectively. This paper proposes a high-performance detection model for cigarette defects, named cigarette defect YOLO (CD-YOLO), which builds upon the YOLOv10 network with three major improvements. First, an intra-scale feature interaction (ISFI) module is designed to enhance the model's ability to distinguish different defects. Subsequently, a multi-scale feature fusion (MSFF) network is developed to improve the model's performance in recognizing small-scale and subtle defects. Finally, a lightweight group convolution detection head (LGCDH) is implemented to substantially reduce the model's computational complexity and parameter count, accelerating detection speed. The experimental results demonstrate that the CD-YOLO model achieves a favorable trade-off between accuracy and speed, maintaining a detection speed exceeding 500 FPS, with a mAP@0.5 of 96.2%. Additionally, a novel data augmentation strategy is introduced in this paper, employing low-rank adaptation (LoRA) to fine-tune a pretrained stable diffusion model, which generates synthetic defect samples to alleviate data scarcity.

Keywords: cigarette appearance defects; defect detection; YOLOv10; deep learning; defect generation

1. Introduction

The tobacco industry plays a crucial role in China's economy, contributing significantly to

national revenue. Among tobacco products, cigarettes account for the majority of production, with annual output exceeding 200 billion units. During high-speed, large-scale manufacturing, various appearance defects may arise due to equipment or raw material flaws, such as dotted, wrinkled, misaligned, and unfiltered [1]. These defects are the most visually apparent indicators of cigarette quality, directly affecting consumer perception and brand reputation [2]. Consequently, tobacco companies must implement stringent quality control measures to prevent defective cigarettes from entering the market.

Currently, the tobacco industry primarily relies on infrared photoelectric detection and traditional machine vision methods for cigarette appearance defect inspection. Infrared photoelectric detection offers advantages, such as noncontact operation and high detection speed, but it struggles to classify defect types [3]. Traditional machine vision methods can identify defects, but they require handcrafted feature extraction [4]. Both approaches fall short of meeting the demands of modern production lines, so there is an urgent need to develop more efficient methods for detecting cigarette appearance defects.

With the rapid development of deep learning, the use of deep learning for defect detection has sparked the interest of researchers [1,2,5–7]. Compared to traditional detection methods, deep learning eliminates the need for manual feature extraction and can accurately identify the location and type of defects. Yuan et al. [5] improved the YOLOv4 model by introducing a channel attention mechanism and optimizing the selection of clustering centers using the K-means++ algorithm, achieving 91.77% mAP@0.5 at 18.8ms inference speed. Liu et al. [1] proposed a detection model for cigarette appearance defects based on C-CenterNet, using ResNet50 as the backbone feature extraction network. By incorporating the convolutional block attention module (CBAM) and deformable convolution, their model achieved a mAP@0.5 of 95.01% with an inference speed of 8.9ms. Wu et al. [6] proposed a detection model for cigarette appearance defects based on an enhanced SSD framework, combining variational Bayesian inference, BIoU loss function, and an improved activation function. Their model increased the mAP@0.5 by 1.2% and reduced computational cost by 5.92 GFLOPs.

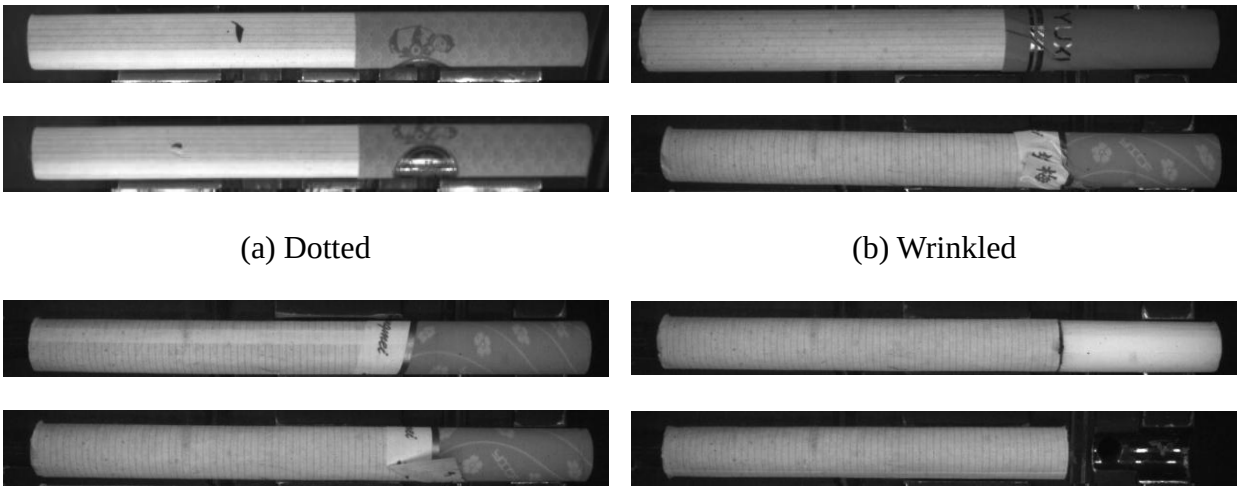
Although these studies have achieved promising results, there remains room for improvement in detection accuracy and speed. Moreover, the high computational cost and large parameter count hinder model deployment on edge devices. Therefore, designing more efficient detection algorithms for cigarette appearance defects is necessary to achieve high accuracy and speed while reducing model complexity.

This paper proposes a high-performance detection model for cigarette appearance defects based on the YOLOv10 [8] network, cigarette defect YOLO (CD-YOLO). First, an intra-scale feature interaction (ISFI) module was designed and applied to the feature extraction network's final feature layer, enhancing the model's global representation capabilities and the discriminative ability for different cigarette appearance defects. Second, a multi-scale feature fusion (MSFF) network was developed, effectively integrating shallow spatial and deep semantic features to reduce the omission rates of minor defects. Third, a lightweight group convolution detection head (LGCDH) is implemented to substantially reduce the model's computational complexity and parameter count, thereby accelerating detection speed. Additionally, to mitigate the shortage of defect samples, a low-rank adaptation (LoRA) technique [9] is employed to fine-tune the stable diffusion model [10] to generate high-quality cigarette appearance defect images. Experimental results demonstrate the effectiveness of the proposed approach, which exhibits a favorable trade-off between accuracy and speed in detecting cigarette appearance defects.

2. Dataset construction and augmentation

2.1. Cigarette appearance defects dataset

Common cigarette appearance defects can be classified into four categories (dotted, wrinkled, misaligned, and unfiltered [1,2,5,6]), as shown in Figure 1. Dotted defects are characterized by irregular black spots on the cigarette surface, typically caused by holes from tobacco stems or stains on the cigarette paper. Wrinkled defects are identified by cigarette surface folds or gaps and bulges at the tipping paper joints. Misaligned defects are manifested through warping or misplacement of the tipping paper seams. Unfiltered defects are distinguished by the absence of the filter rod or missing tipping paper wrapping around the filter. All cigarette image samples for this study were obtained from the production lines of China Tobacco Yunnan Industrial Co., Ltd.



(a) Dotted (b) Wrinkled

(c) Misaligned (d) Unfiltered

Figure 1. Cigarette examples with appearance defects.

2.2. Data augmentation

Due to the limited quantity of defect samples available on the production line, this study employs text-to-image generation based on stable diffusion [10] to synthesize defect samples. As one of the most powerful generative models today, stable diffusion has gained significant attention in various fields due to its strong interpretability and outstanding generative capability. However, the original stable diffusion model cannot directly generate images of cigarette appearance defects. Therefore, this study leverages LoRA [9] to fine-tune stable diffusion for generating defects, as illustrated in Figure 2.

To enable stable diffusion to generate specific categories of cigarette appearance defects, this study trains individual LoRA models $\Delta\theta_i (i=1,2,3,4)$ for each defect category using its corresponding defect images $\mathcal{J}_i (i=1,2,3,4)$ on the UNet ϵ_θ [11]. The loss function for training LoRA is defined as follows:

$$\mathcal{L}(\Delta\theta_i) = E_{\epsilon,\mathrm{t}}[\| \epsilon - \epsilon_{\theta+\Delta\theta_i}(z_{ti}, t, c_i) \|^2] \quad (1)$$

In the formula, $\epsilon \sim N(0, I)$ represents Gaussian noise randomly sampled from a standard normal distribution, z_{ti} denotes the noise latent embedding corresponding to the input image $\mathcal{J}_i$ at

timestep t, and c_i is the text embedding encoded from the text prompt p_i. $\epsilon_{\theta+\Delta\theta_i}$ denotes the UNet integrated with the LoRA model. By minimizing the loss function $\mathcal{L}(\Delta\theta_i)$, the model optimizes the parameters $\Delta\theta_i$ for updates. Once fine-tuning is complete, $\Delta\theta_i$ is fixed and stored. Afterward, the UNet $\epsilon_{\theta+\Delta\theta_i}$ integrated with $\Delta\theta_i$ can be used as the noise prediction network in the denoising step. By inputting text prompts corresponding to different defect categories, the model can generate specific cigarette appearance defect image samples.

This study also designs a series of image-processing-based data augmentation methods tailored to the actual characteristics of cigarette appearance defect images. The data augmentation techniques used include GaussNoise [13], ColorJitter [14], Affine [15], RandomFlip [16], and RandomCrop [17]. These methods effectively simulate defects under complex real-world scenarios, enhancing the model's robustness.

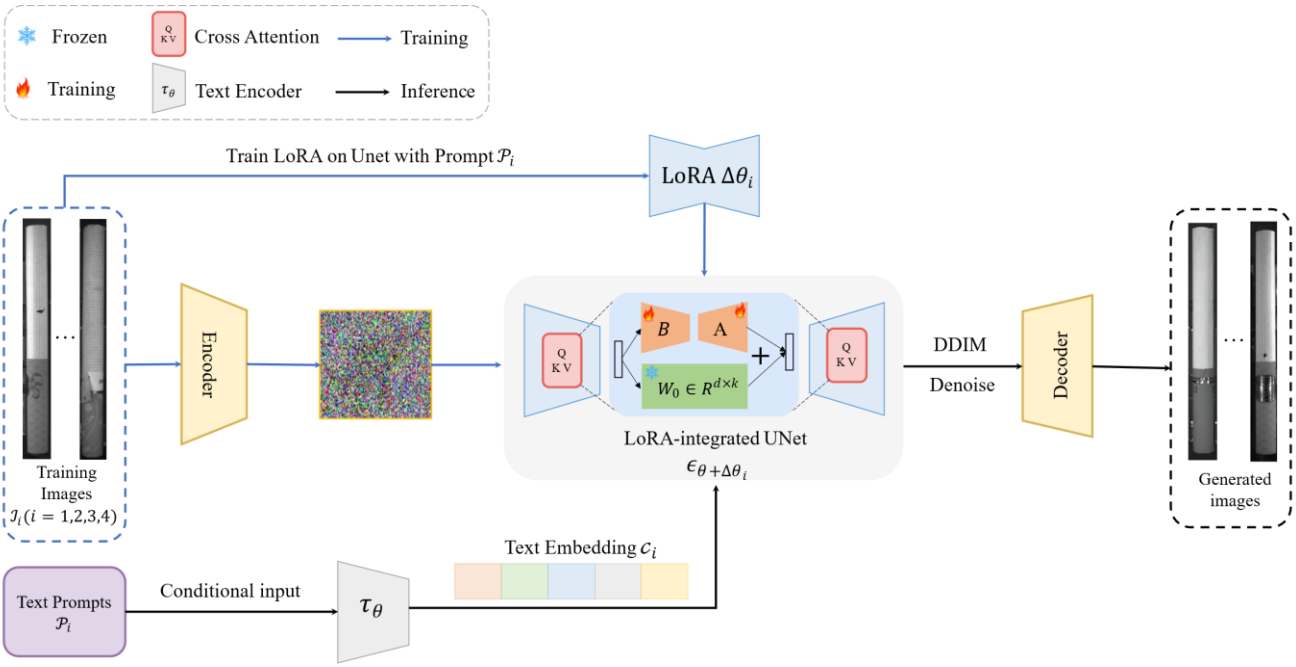


Figure 2. Data generation pipeline. For each category of cigarette appearance defect images $\mathcal{I}_i (i=1,2,3,4)$, separate LoRA models $\Delta\theta_i$ are trained individually. These trained LoRA models are then integrated into the UNet ϵ_θ. The resulting UNet $\epsilon_{\theta+\Delta\theta_i}$, incorporating the $\Delta\theta_i$ parameters, is used as the noise prediction network during the denoising step. Finally, specific cigarette appearance defect image can be generated by providing corresponding text prompts p_i for each defect category.

2.3. Dataset partitioning

The original dataset for cigarette appearance defects comprises 2,135 cigarette images, each with a resolution of 1200 × 146 pixels. Defects are annotated into four categories: dotted, wrinkled, misaligned, and unfiltered. The dataset is partitioned into training, validation, and test sets in a ratio of 6:2:2. Following data expansion, the training set is expanded to 10,266 images. It is important to note that the expanded images are used only for training; all testing and validation are conducted using only real-world defect samples. Table 1 presents the number of images for each defect type in the dataset.

Table 1. The original dataset and the expanded dataset.

Defect types		Dotted	Unfiltered	Wrinkled	Misaligned
Original dataset	Train.	371	316	180	57
	Val.	125	105	59	20
	Test.	126	106	59	20
Expanded dataset	Train.	2618	2538	2566	2554
	Val.	125	105	59	20
	Test.	126	106	59	20

3. Methodology

3.1. Model overview

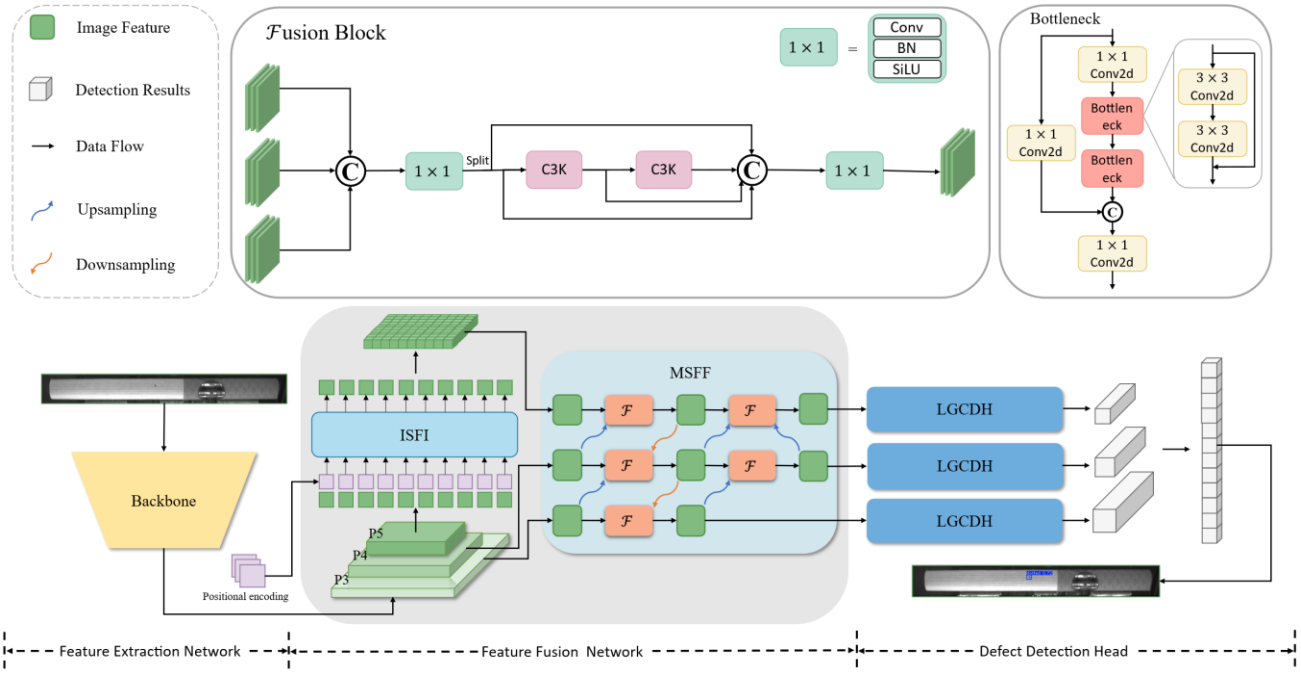


Figure 3. Overview of CD-YOLO. The input images are fed into the feature extraction network, where a series of convolutional and pooling operations transform the image into multi-scale feature maps. The multi-scale feature maps from the last three stages {P3, P4, P5} are then passed to the feature fusion network, which performs comprehensive global modeling and feature fusion through the ISFI module and the MSFF network. Finally, the enhanced features are sent to the defect detection, which predicts the target categories and locations.

The CD-YOLO model consists of a feature extraction network, a feature fusion network, and a defect detection head. Specifically, the feature extraction network captures multi-scale semantic and texture features from the input cigarette images and passes the final three feature maps {P3, P4, P5} to the feature fusion network. Subsequently, the efficient ISFI module performs cross-scale feature interaction, and the MSFF network effectively integrates shallow spatial details with deep semantic information, yielding a set of high-quality fused features. Finally, the LGCDH generates the defect

category, bounding box coordinates, and confidence score based on the fused features. Figure 3 presents an overview of CD-YOLO. The structural details of ISFI, MSFF, and LGCDH are elaborated in Sections 3.2 to 3.4.

3.2. Intra-scale feature interaction module

The original YOLOv10 model [8] employs a convolutional neural network (CNN) for feature extraction, which is effective at capturing local patterns but limited in its ability to model global contextual information. Consequently, the model struggles to distinguish between different defect types and to separate defects from the background, resulting in a higher risk of false detections. The study designs a self-attention-based ISFI module, as shown in Figure 4, to enhance the model's discrimination of different cigarette appearance defects. Self-attention [18] is widely used in various visual tasks due to its remarkable global modeling capability [19-20]. However, it exhibits high computational complexity and memory footprint [8]. This is primarily due to the dynamic statistics computation of LayerNorm [21] within the attention mechanism, which introduces additional computational overhead, thus significantly hindering the running speed [22].

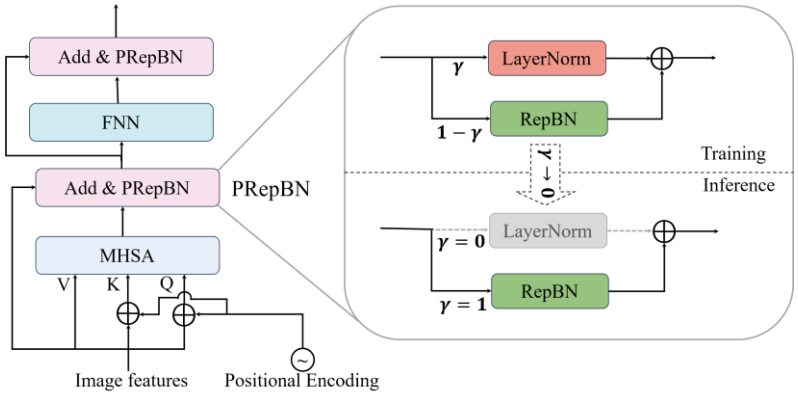


Figure 4. The structure of ISFI. ISFI adopts a progressive re-parameterized BatchNorm strategy [22], where LayerNorm is gradually replaced by RepBN during training. At the beginning of training, $\gamma = 1$, and LayerNorm dominates. By the end of training, $\gamma = 0$, and BatchNorm completely replaces LayerNorm.

To address this, this study explores replacing LayerNorm with BatchNorm [23] to accelerate inference. Directly leveraging BatchNorm tends to lead to unsatisfactory performance. Hence, this study considers a progressive strategy [22] to replace LayerNorm with BatchNorm gradually. At the early stage of training, LayerNorm is retained to ensure stable optimization, as it normalizes features independently of batch statistics and thus effectively alleviates gradient instability. As training progresses, BatchNorm is gradually introduced, leveraging more reliable batch-level statistics to enhance feature discrimination. By the end of training, the model transitions to a pure BatchNorm form, eliminating the need for per-sample statistics during inference and thereby significantly reducing computational overhead. This process can be formulated as follows:

$$\mathrm{PrepBN}(X) = \gamma \mathrm{LN}(X) + (1-\gamma)\mathrm{RepBN}(X) \tag{2}$$

$$\mathrm{RepBN}(X) = \mathrm{BN}(X) + \eta X \tag{3}$$

where γ is a hyperparameter used to control the output ratio of the two normalization layers and

η is a learnable hyperparameter and is jointly trained in an end-to-end manner. At the early stages of training, $\gamma = 1$, making LayerNorm the dominant component. As training progresses, γ is progressively reduced to 0, enabling RepBN to replace LayerNorm by the end of training fully. The decay strategy of γ is defined as follows:

$$\gamma = \frac{T - T_{cur}}{T}, \gamma \in [0,1] \tag{4}$$

where T is the total number of training steps and T_{cur} is the current step. This linear decay strategy ensures a smooth transition from LayerNorm to BatchNorm. By adopting this strategy, the global modeling capability of self-attention can be integrated into the model with lower computational cost, enhancing the model's discrimination of different cigarette appearance defects without affecting its inference speed.

3.3. Multi-scale feature fusion network

The original YOLOv10 [8] model employs a modified PANet [24] structure for feature fusion. However, when applied to the task of detecting cigarette appearance defects, the model exhibits suboptimal performance in detecting minor and subtle defects. Upon revisiting the feature fusion network in YOLOv10, this study observes that each feature layer only interacts with its adjacent layers, failing to balance information across all feature levels effectively. As a result, shallow detail features are progressively diluted during transmission, leading to poor detection capability for small targets.

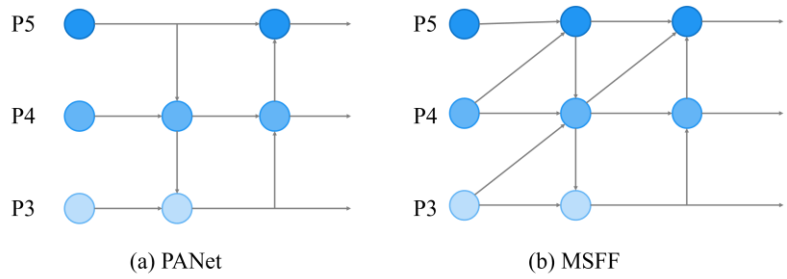


Figure 5. Feature fusion network evolution design from P3 to P5. (a) PANet used in YOLOv10 enhances features through both top-down and bottom-up pathways. (b) MSFF network builds upon PANet by incorporating cross-layer connections from lower network layers.

To address this issue, this paper presents an efficient MSFF network based on the PANet architecture, as shown in Figure 5. Inspired by the RepGFPN structure in DAMO-YOLO [25], this study introduces additional cross-layer connections at each feature level in their design. This allows every feature node to comprehensively integrate multi-scale information, mitigating the information degradation caused by long transmission paths in the original structure. Such cross-layer feature fusion is particularly critical in cigarette defect detection, where defect sizes can vary by up to a factor of 500. Preserving fine spatial details is crucial for enhancing the model's ability to detect small defects and complex textures.

3.4. Lightweight group convolution detection head

The detection head in YOLOv10 accounts for nearly one-fourth of the total model parameters. It adopts a decoupled detection head, separating classification and regression into two independent branches [34]. Although this design substantially improves detection accuracy in scenarios with numerous categories, it significantly increases the model's parameter count [26]. Given that the cigarette appearance defect detection task involves only four defect categories, such an approach offers diminishing returns, as the accuracy gains do not justify the considerable increase in model complexity. To this end, this study designs a LGCDH, as illustrated in Figure 6.

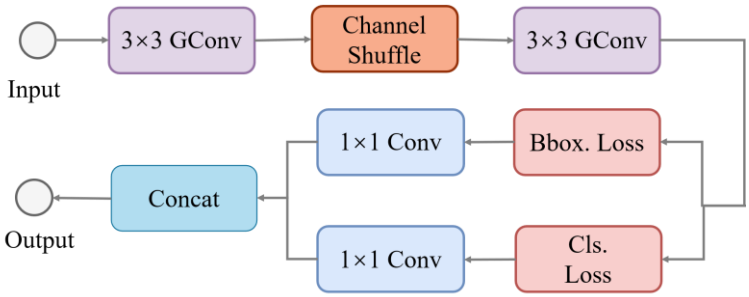


Figure 6. Structure of LGCDH. Input features pass through two 3×3 grouped convolutions for feature extraction, followed by separate 1×1 convolutions for output mapping. The channel shuffle operation is employed to mitigate the information isolation caused by consecutive grouped convolutions.

This study integrates the classification and regression heads by enabling them to share two 3×3 group convolutions [27] for feature extraction, followed by separate 1×1 convolutions for output mapping. Grouped convolutions can significantly reduce the model's parameter count and computational complexity [28]. Considering that each group contains only incomplete and partial representations of the graph, sufficient feature capture is unlikely without communication among multiple group convolutions. This study introduces a channel shuffle operation [29] after the first grouped convolution to enable cross-group information flow and effectively mitigate the feature isolation caused by grouping.

4. Experiments

4.1. Experimental setup

The experiments in this study were implemented using the PyTorch framework in Python. During training, the input image resolution was set to 640×640 with a batch size of 16. Stochastic gradient descent (SGD) with a cosine annealing learning rate schedule was adopted as the optimizer. The initial learning rate was set to 0.01, with a warm-up period of 3 cycles, and the model was trained for 300 epochs. All experiments were performed on a high-performance workstation equipped with an NVIDIA GeForce RTX 4090 (24GB) GPU, an Intel Core i5-13400F CPU, 32GB of RAM, and running Ubuntu 22.04 LTS. Both training and testing were conducted on the same hardware and software environment to ensure consistency.

4.2. Evaluation metrics

In this study, mAP@0.5 and mAP@0.5:0.95 were used as evaluation metrics to measure model accuracy. Specifically, mAP@0.5 denotes the mean average precision (AP) across all categories at an intersection over union (IoU) threshold of 0.5, and mAP@0.5:0.95 is computed by averaging AP over 10 IoU thresholds ranging from 0.5 to 0.95 in increments of 0.05.

To evaluate model complexity, this study adopts two metrics: the number of parameters (Params) and floating-point operations (FLOPs). Params refer to the total count of learnable weights in the model, reflecting its spatial complexity. FLOPs indicate the total number of floating-point operations required for a single forward pass, serving as a measure of computational complexity. Additionally, frames per second (FPS) is employed to evaluate detection speed, representing the number of images the model can process per second.

4.3. Results and analysis

Figure 7 compares the precision–recall (PR) curves of the baseline YOLOv10-N and the improved CD-YOLO on different cigarette appearance defect categories. The baseline model achieves satisfactory performance on relatively simple defects, such as dotted (93.8%) and unfiltered (99.5%), but struggles with more challenging defects, such as wrinkled (72.2%) and misaligned (85.4%), where the PR curves show lower precision at high recall levels. In contrast, CD-YOLO significantly enhances the detection of these challenging defects, with the AP@0.5 of wrinkled and misaligned increasing to 93.6% and 96.8%, respectively. The overall mAP@0.5 improves from 87.7% to 96.2%. These results indicate that the proposed model effectively reduces false and missed detections, particularly for subtle and small-scale defects, leading to more stable PR curves across categories.

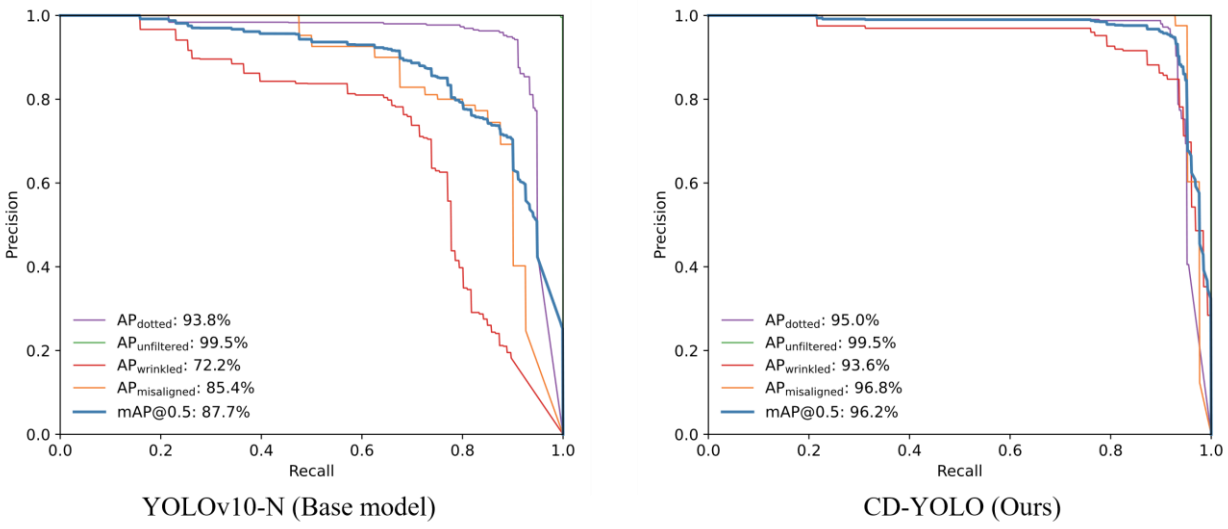
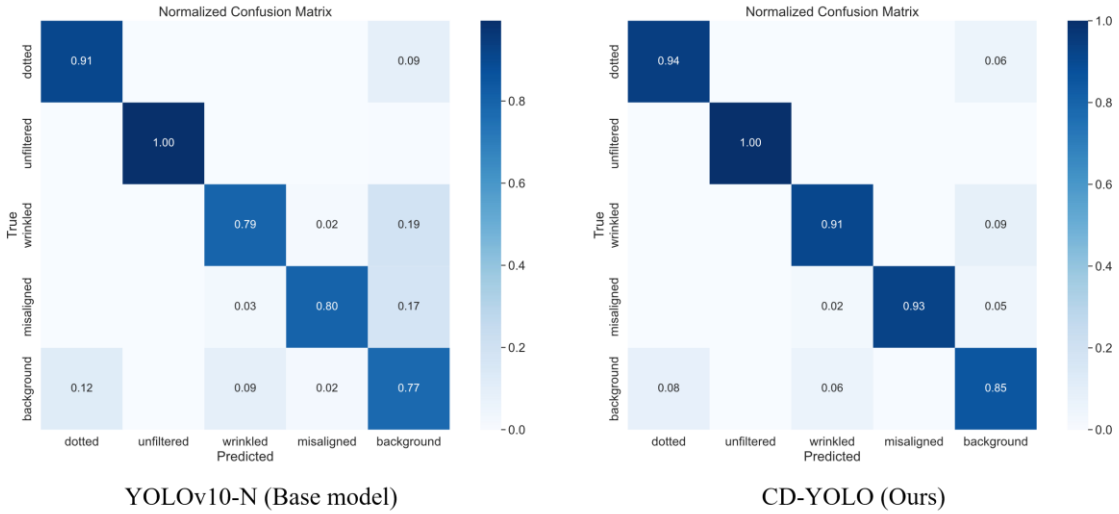


Figure 7. PR curves comparison between the baseline YOLOv10-N and the proposed CD-YOLO on cigarette appearance defect detection. CD-YOLO achieves a higher area under the PR curve compared with YOLOv10-N, indicating that it maintains superior precision across a wide range of recall levels.

The confusion matrix in Figure 8 offers a more detailed assessment of classification performance across defect categories. Compared with the baseline model, CD-YOLO achieves lower misclassification rates, particularly reducing false recognition between visually similar defects such as wrinkled and misaligned. Moreover, the model exhibits stronger discriminative ability across different defect types, leading to fewer missed detections and an enhanced sensitivity to subtle or small-scale defects. These results demonstrate that the proposed improvements enhance the model's ability to discriminate between different defect types and increase its sensitivity to subtle defects.



YOLOv10-N (Base model) CD-YOLO (Ours)

Figure 8. Normalized confusion matrices of the baseline YOLOv10-N (left) and the improved CD-YOLO (right) on cigarette appearance defect detection. CD-YOLO shows reduced misclassification rates, particularly for visually similar and subtle defects, such as wrinkled, misaligned, and dotted, indicating improved discriminative ability and sensitivity.

4.4. Comparison with different mainstream models

Table 2 presents the experimental comparison between the proposed CD-YOLO model and several mainstream real-time object detectors. All comparisons are conducted on the expanded training dataset to ensure a fair and comprehensive evaluation. Compared with the baseline model (YOLOv10-N [8]), CD-YOLO reduces the number of parameters and computational complexity by 4.3% and 14.9%, respectively. At the same time, it achieves gains of 2.7% in mAP@0.5 and 2.1% in mAP@0.5:0.95, while boosting FPS by 10.8%. To further illustrate the detection performance, Figure 9 depicts the Precision–Recall (PR) curves of different models. As shown, CD-YOLO consistently maintains higher precision across a wide range of recall values, indicating its superior ability to balance false positives and false negatives. These results demonstrate that the proposed model achieves the favorable trade-off between efficiency and accuracy.

Table 2. Comparison with different mainstream models. Our CD-YOLO achieves a reasonable trade-off between speed and accuracy, with further reductions in both the number of parameters and computational complexity. To ensure fairness in the experiments, all models employed the same data augmentation strategy. b = 4 denotes a batch size of 4 during inference. The bold values represent the best results.

Model	AP@0.5(%) & AP@0.5:0.95(%)				map@0.5(%)	map@0.5:0.95(%)	#Param.(M)	FLOPs (G)	FPS(b=4)
	dotted	unfiltered	wrinkled	misaligned					
YOLOv7-tiny [26]	94.6/43.8	99.5/93.2	80.7/33.9	93.1/53.8	91.9	56.2	6.0	13.2	384
DAMO-YOLO-T [25]	94.5/46.2	99.5/93.8	80.2/37.5	88.6/52.6	90.7	57.5	8.5	18.1	111
Gold-YOLO-N [30]	94.8/45.3	99.5/93.1	83.5/36.2	90.3/52.5	92.0	56.7	5.6	12.1	329
RT-DETR-R18 [31]	93.8/45.7	99.5/93.7	78.8/32.7	78.3/50.8	87.6	55.7	20.0	60.0	217
YOLOv8-N [12]	95.1/48.0	99.5/94.5	84.2/37.7	91.8/59.3	92.6	59.9	3.2	8.7	**581**
YOLOv9-T [32]	93.7/47.0	99.5/**94.8**	83.2/39.9	88.2/55.3	91.2	59.3	**2.0**	7.6	299
YOLOv10-N [8]	**95.2**/48.5	99.5/93.2	89.6/45.3	90.8/59.2	93.8	61.5	2.3	6.7	483
YOLO11-N [33]	94.5/46.6	99.5/94.1	85.9/39.0	89.0/55.7	92.2	58.8	2.6	6.5	563
CD-YOLO(Ours)	95.0/**49.0**	**99.5**/93.8	**93.6**/**47.0**	**96.8**/**64.1**	**96.2**	**63.5**	2.2	**5.7**	535

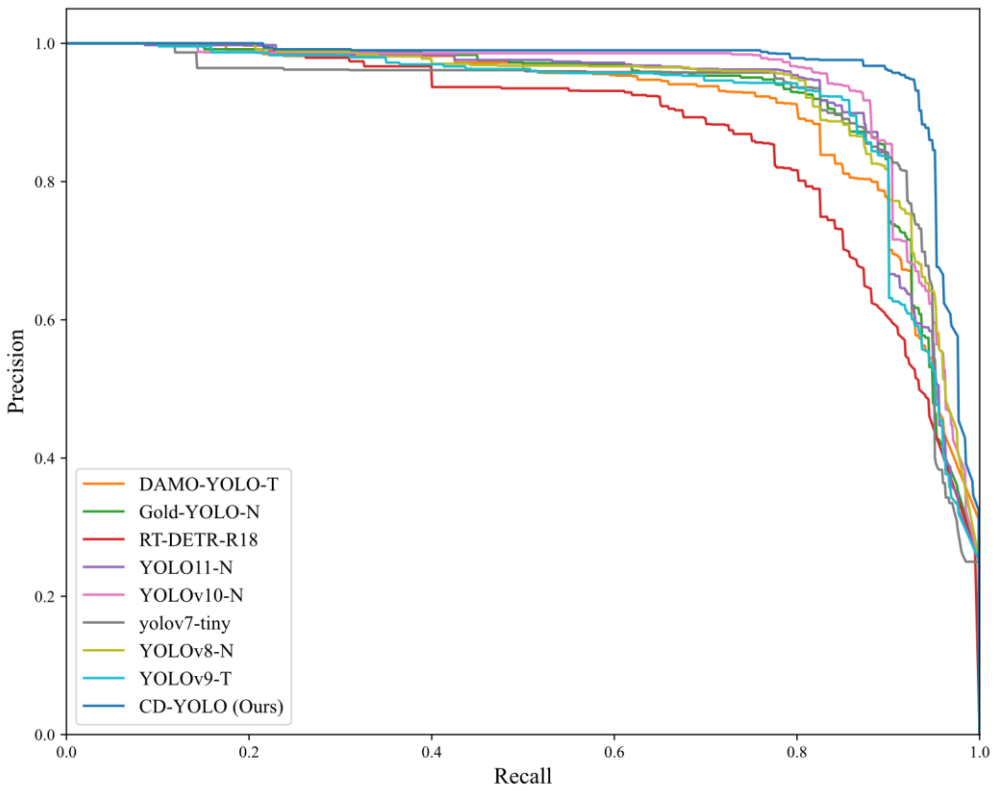


Figure 9. PR curves of different real-time object detection models on the cigarette appearance defect dataset. The proposed CD-YOLO achieves consistently better precision across varying recall levels compared with baseline methods.

4.5. Methodology analysis

4.5.1. Analysis of model improvement effectiveness

This paper conducted comparative experimental analyses on the proposed improvements. Multiple improved models were constructed sequentially, and the results were compared using the same test data. Table 3 shows the ablation study results. The dataset augmentation strategy yielded improvements of 6.2% in mAP@0.5 and 5.3% in mAP@0.5:0.95. Implementation of the MSFF network further increased mAP@0.5 by 1.2% and mAP@0.5:0.95 by 1.4%. Subsequent integration of the ISFI module provided additional gains of 0.9% in mAP@0.5 and 0.3% in mAP@0.5:0.95. All combined improvements resulted in peak performance metrics of 96.2% mAP@0.5 and 63.5% mAP@0.5:0.95. Concurrently, model parameters and FLOPs were reduced to 2.2M and 5.7G, respectively. Overall, the improved model not only reduces computational complexity and parameter count but also significantly boosts detection accuracy and speed.

Table 3. Results of the ablation study on model improvement. Bold values denote the best value.

Data augmentation	MSFF	ISFI	LGCDH	AP@0.5(%) & AP@0.5:0.95(%)				#Param.(M)	FLOPs (G)	map@0.5(%)	map@0.5:0.95(%)
				dotted	unfiltered	wrinkled	misaligned				
				93.8/46.9	99.5/93.3	72.2/29.6	85.2/53.3	2.3	6.5	87.7	55.7
√				95.2/48.5	99.5/93.2	89.6/45.3	90.8/59.2	2.3	6.5	93.8	61.5
	√			**95.7**/49.3	99.5/94.4	81.9/38.4	90.7/57.8	2.6	6.5	91.9	59.7
		√		94.4/47.9	99.5/94.3	79.1/36.7	88.7/55.5	2.1	6.3	90.4	58.6
			√	93.4/46.8	99.5/94.0	76.1/36.0	87.2/56.1	2.0	5.9	89.1	58.2
	√	√		94.7/**49.8**	99.5/**94.5**	85.0/38.5	90.8/58.5	2.5	6.4	92.5	60.3
	√		√	95.5/49.5	99.5/94.0	84.0/37.8	90.1/58.2	2.3	5.8	92.3	59.9
		√	√	94.8/48.2	99.5/94.1	80.5/38.3	90.1/55.7	**1.9**	5.7	91.2	59.1
	√	√	√	94.5/50.2	99.5/**94.5**	84.3/38.6	92.5/58.7	2.2	5.7	92.7	60.5
√	√	√	√	95.0/49.0	**99.5**/93.8	**93.6**/47.0	**96.8**/64.1	2.2	**5.7**	**96.2**	**63.5**

4.5.2. Analyses for data augmentation

To address the limitations of insufficient sample size and class imbalance in the original dataset, this paper employed LoRA to fine-tune the stable diffusion model for generating realistic cigarette appearance defects. This approach enables the synthesis of defect samples that closely resemble real defects, thereby enriching the diversity of the training data. As illustrated in Figure 10, the generated samples exhibit high visual similarity to real defects, which provides a valuable complement to the limited real-world data and enhances the effectiveness of subsequent model training.

To quantitatively evaluate the distributional differences between the generated data and the real data, this paper performed a t-distributed stochastic neighbor embedding (t-SNE) visualization. Specifically, deep features were extracted from both real and generated image sets using a ResNet50 network pretrained on ImageNet. The extracted high-dimensional features were then projected into a two-dimensional space using t-SNE for visualization. To ensure feature comparability, the final fully connected layer of ResNet50 was removed, retaining only the convolutional feature representations, and all input images were normalized with the same preprocessing procedure. As shown in Figure 11,

although a certain gap still exists between the distributions of the generated and real data, the generated samples approximate the real data distribution reasonably well.

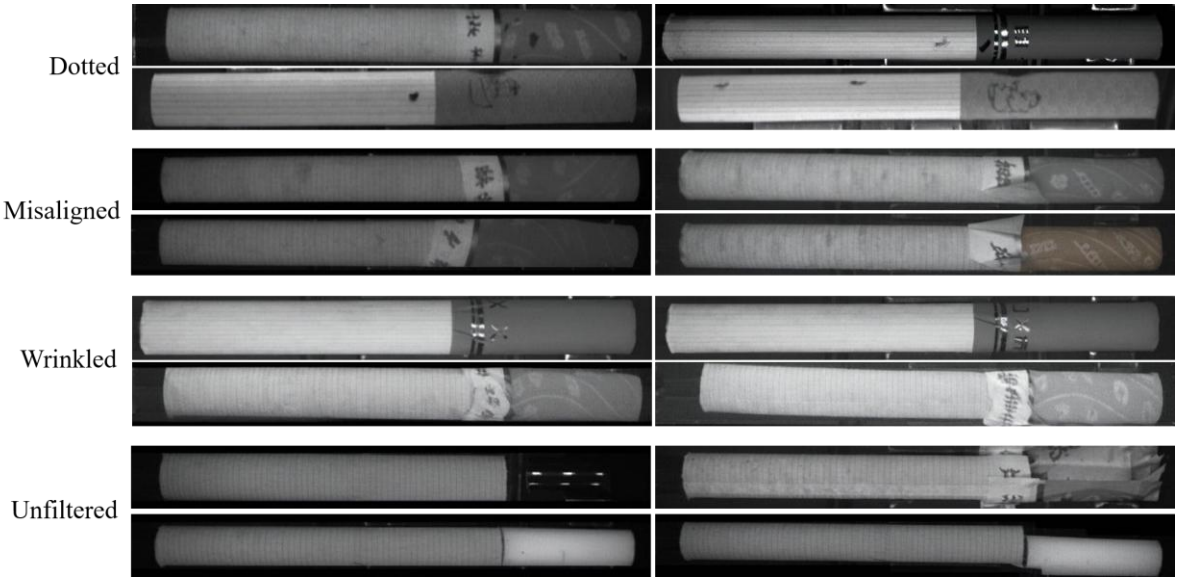


Figure 10. Examples of cigarette appearance defect samples generated using LoRA-tuned stable diffusion. The synthesized defects exhibit visually realistic characteristics.

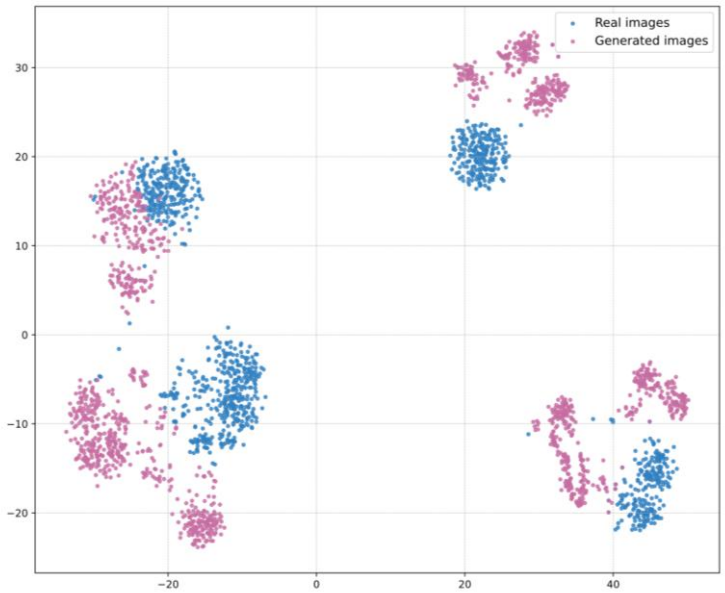


Figure 11. t-SNE visualization of feature distributions for real and generated images. A closer overlap between the two distributions indicates that the generated defect samples better approximate the characteristics of real defects. Although there are some differences between the generated and real data in feature distribution, the generated samples still closely approximate the real data distribution.

In addition, we designed a series of image processing–based data augmentation methods to expand the dataset, reduce the risk of overfitting, and improve the robustness and generalization of the model (see Figure 12). GaussNoise introduces normally distributed noise to enhance the model's adaptability to uncertainty and noisy conditions; ColorJitter randomly adjusts color attributes, enabling the model to cope with variations in illumination and color; Affine applies linear transformations to strengthen the recognition of defects under different perspectives, scales, and positions; RandomCrop crops images randomly, improving the model's ability to detect small defects; and RandomFlip performs horizontal or vertical flips to increase sample diversity. These augmentation strategies effectively simulate the diversity of data encountered in complex industrial environments.

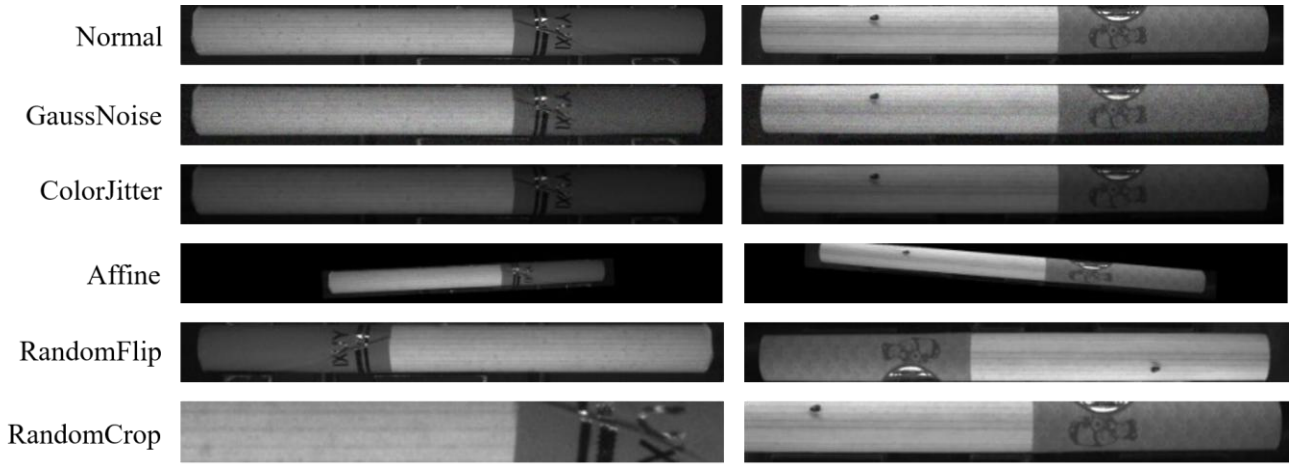


Figure 12. Five image processing-based data augmentation methods for cigarette appearance defect images.

Table 4 reports the number of expanded samples for each type of cigarette appearance defect as well as the detection performance of the model under different data augmentation strategies. When only image processing-based augmentation is applied, the model achieves improvements of 0.7% in mAP@0.5 and 0.6% in mAP@0.5:0.9. Training solely with generated data results in lower detection accuracy compared with real data. However, when generated data is combined with the original dataset, the model's performance increases substantially, with gains of 3.9% in mAP@0.5 and 2.7% in mAP@0.5:0.95. This improvement can be attributed to the novel features introduced by generated data, which supplement rare classes and previously unseen characteristics, thereby strengthening model generalization. Furthermore, when the original dataset is first expanded with generated data and subsequently augmented through image processing, the proposed model achieves the best performance, yielding improvements of 6.1% in mAP@0.5 and 5.8% in mAP@0.5:0.95. These findings demonstrate that dataset expansion alone can substantially enhance detection accuracy without introducing additional computational overhead.

Table 4. Performance comparison of the model trained with different data compositions. The augmented data are obtained through image processing–based augmentation on the original data, while the generated data are produced using stable diffusion with LoRA fine-tuning. The bold values represent the best results.

Original data	Augmented Data	Generated Data	Dataset Size				Test	
			dotted	unfiltered	wrinkled	misaligned	mAP@0.5(%)	mAP@0.5:0.95(%)
√			371	316	180	57	87.7	55.7
	√		1500	1276	724	228	88.3	56.8
		√	413	311	517	591	72.4	43.7
	√	√	1913	1587	1241	819	92.5	59.4
√	√		1871	1592	904	285	89.2	56.7
√		√	784	627	697	648	91.5	58.9
√	√	√	2618	2538	2566	2554	**93.8**	**61.5**

Note. When combining all three types of data, the original and generated samples were first merged, and data augmentation was subsequently applied. Therefore, the dataset size in the last row does not follow a simple accumulation pattern of the previous rows.

4.6. Performance evaluation across different resolutions

Due to the fine-grained nature of defects, input resolution has a notable impact on detection performance. To assess the scalability and efficiency of the model under different resolutions, we conducted experiments at multiple input sizes, as shown in Table 5. The results reveal a clear trade-off: higher resolutions improve accuracy but reduce inference speed. Specifically, both mAP@0.5 and mAP@0.5:0.95 increase as resolution rises, demonstrating that finer details are better captured at larger input sizes. However, the computational cost also grows significantly, FPS drops from 959 at 224×224 to 254 at 1024×1024. This suggests that low resolutions (224×224 and 256×256) are suitable for real-time scenarios where speed is essential, whereas higher resolutions (512×512 or 640×640) are more appropriate when accuracy is the priority. Overall, resolutions between 512×512 and 640×640 offer the best balance, providing substantial accuracy gains without excessive loss of speed.

Table 5. Performance evaluation of the CD-YOLO at different input resolutions. b = 4 denotes a batch size of 4 during inference. The bold values represent the best results.

Input resolution	**map@0.5(%)**	**map@0.5:0.95(%)**	**FPS(b=4)**
224×224	49.6	30.2	959
256×256	67.2	38.9	835
416×416	93.8	57.4	698
512×512	95.7	61.1	586
640×640	**96.2**	**63.5**	535
1024×1024	94.6	54.8	**254**

4.7. Visualization of detection results

Figure 13 compares the effective receptive fields (ERF) [35] of CD-YOLO and the baseline model during feature extraction. The visualization clearly illustrates that incorporating the ISFI module significantly expands CD-YOLO's effective receptive field. This enhancement strengthens the model's global feature modelling capability, allowing it to capture the overall structure of the cigarette better and improve the distinction between defect regions and the background, thereby effectively reducing false detections.

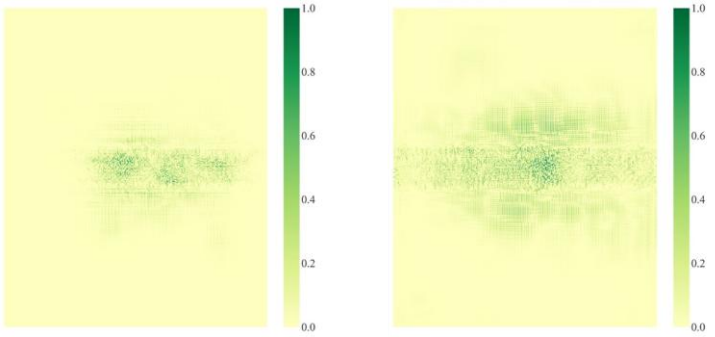


Figure 13. The ERF of YOLOv10-N and CD-YOLO in the last stage of feature extraction in the backbone. A more widely distributed dark area indicates a larger ERF.

Figure 14 compares the Grad-CAM maps [36] of CD-YOLO and the baseline model within the feature fusion network. It can be observed that although the baseline model attends to defect areas, it also responds to non-defect regions and fails to highlight small targets and complex texture defects effectively. In contrast, after the redesign of the feature fusion network, CD-YOLO more precisely captures defect features and generates stronger responses at defect locations.

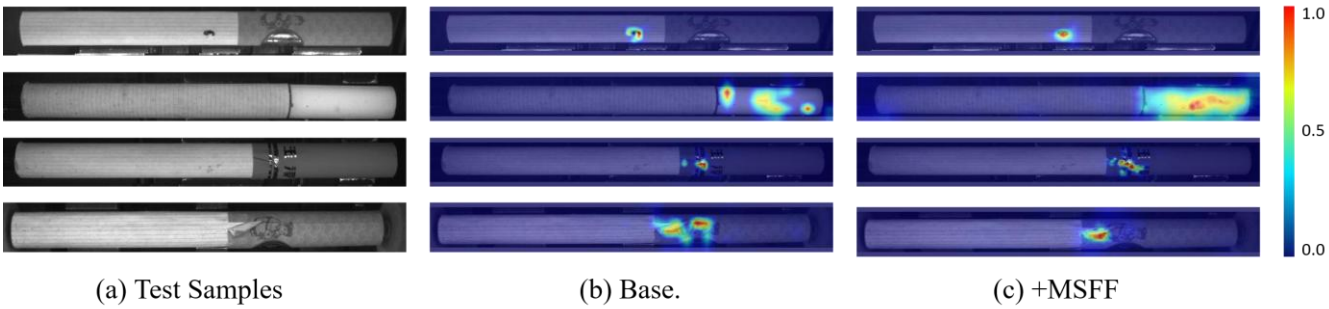


Figure 14. Visualization of heatmaps at the feature fusion networks of the YOLOv10-N and CD-YOLO, where warmer colors indicate higher attention to the corresponding regions.

Figure 15 compares the detection results of CD-YOLO and the baseline model on cigarette appearance defect detection. As shown in the figure, the baseline model is prone to confusion between wrinkle defects and misaligned defects, leading to frequent misdetections. In contrast, CD-YOLO, with its enhanced global feature modelling capability, accurately distinguishes between

these defect types. The baseline model also struggles to detect minor and subtle defects, resulting in a high missed detection rate. By leveraging more comprehensive feature fusion, CD-YOLO effectively captures defect regions and significantly reduces missed detections. Furthermore, CD-YOLO not only improves the confidence scores for defect detection but also enhances the ability to detect multiple simultaneous defects.

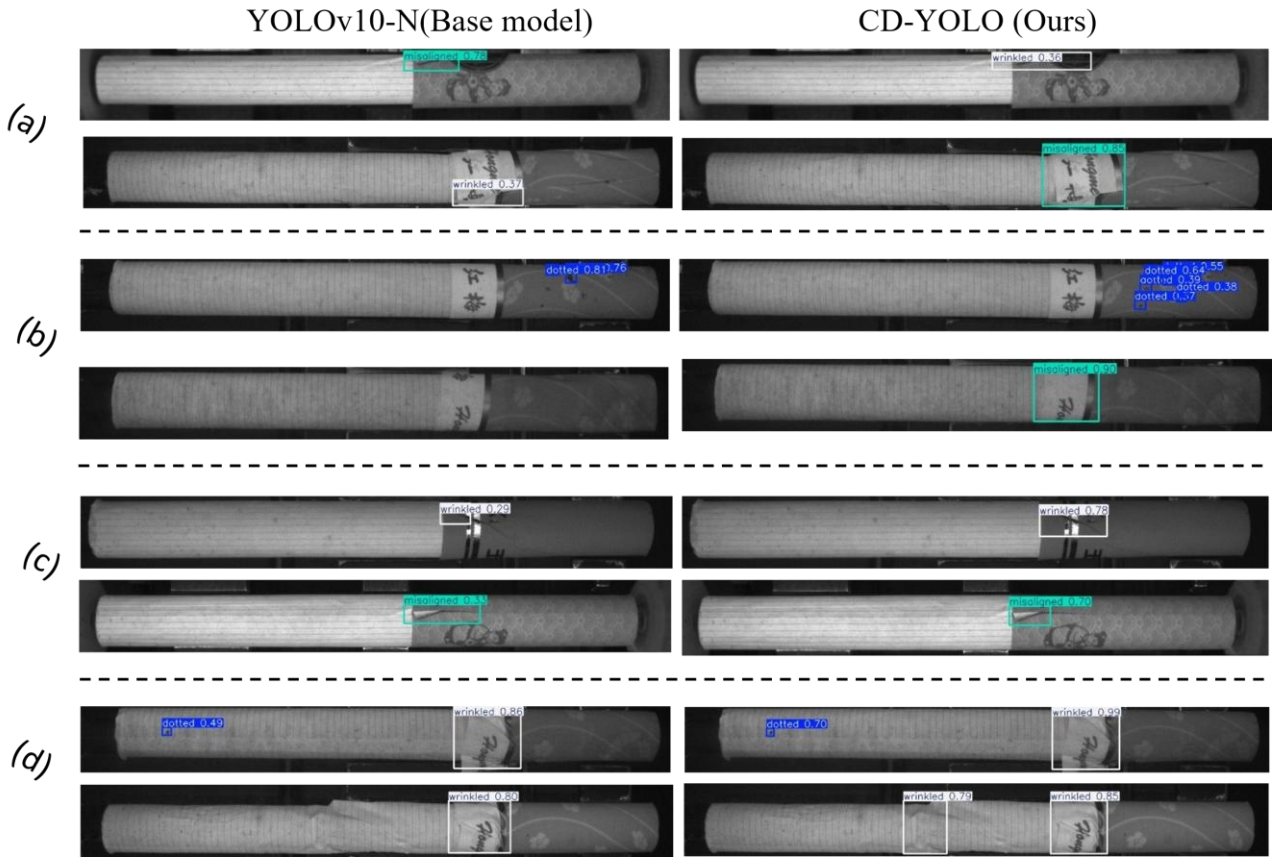


Figure 15. Visualization of the models' predictions: (a) CD-YOLO corrects the false positives produced by the baseline model. (b) CD-YOLO reduces the missed detections of the baseline model. (c) CD-YOLO improves the detection accuracy compared to the baseline model. (d) CD-YOLO accurately detects multiple defects simultaneously.

5. Conclusions and discussion

This paper proposes CD-YOLO, a high-precision and real-time deep learning model for cigarette appearance defect detection. By incorporating the ISFI module, developing the MSFF network, and designing the lightweight detection head, the model achieves an effective balance between detection accuracy and speed. In addition, this study introduces a novel data augmentation strategy that employs LoRA to fine-tune stable diffusion, enabling the generation of synthetic defect samples to alleviate data scarcity. Experimental results demonstrate that CD-YOLO delivers excellent performance in cigarette defect detection tasks, achieving high detection rates while maintaining real-time processing capabilities.

CD-YOLO performs well in identifying defect types seen during training. However, when encountering new or unseen defects on the production line, its detection capability may degrade,

leading to missed detections. In addition, the model heavily relies on high-quality and abundant training data. Although this study leveraged stable diffusion and LoRA techniques to generate synthetic defect samples and mitigate data scarcity, the diversity and authenticity of these generated samples still fall short of fully capturing the complexity of real-world production defects.

Future work may address these issues from two directions. On the one hand, incorporating open-set detection or anomaly detection techniques could improve the model's capability to detect unknown or emerging defects. On the other hand, adopting unsupervised or semi-supervised frameworks may alleviate the dependence on extensive labeled defect datasets.

Author contributions

Yuanyuan Liu developed the conceptual framework, designed the methodology, conducted data analysis, and prepared the initial draft of the manuscript. Guowu Yuan was responsible for data collection, carrying out the literature review, and contributing to manuscript revisions. Hao Wu and Hao Zhou provided project supervision and secured funding for the research.

Use of Generative-AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgements

This work was funded by the Yunnan Provincial Department of Science and Technology - Yunnan University "Double First-Class" Construction Joint Special Key Project (No. 202201BF070001-005); and the Yunnan University Professional Degree Graduate Practice Innovation Fund Project (No. ZC-24249210).

Data availability

The dataset of cigarette appearance defects developed in this study will be publicly released after the manuscript is accepted for publication. It will be made available at https://github.com/lyy8021/CD-YOLO.

Conflict of interest

The authors declare there is no conflict of interest in this paper..

References

1. Liu H, Yuan G, Yang L, Liu K, Zhou H (2022) An appearance defect detection method for cigarettes based on C-CenterNet. *Electronics* 11: 2182. https://doi.org/10.3390/electronics11142182
2. Ding Y, Yuan G, Zhou H, Wu H (2025) ESF-DETR: a real-time and high-precision detection

model for cigarette appearance. *J Real Time Image Process* 22: 54. https://doi.org/10.1007/s11554-025-01632-y
3. Fu J, Zhang J, Wu Z, Xu L, Ye H, Zhang Y (2017) Application of infrared photoelectric detection system to triple filter rod production. *Tob Sci Technol* 50: 85–90. (in Chinese)
4. Li Y, Yang S, Fan L, Xiong Y, Zhu Z, Zhang L (2023) Online inspection of cigarette seam defects based on machine vision. *Tob Sci Technol* 56: 93–98. (in Chinese)
5. Yuan G, Liu J, Liu H, Ma Y, Wu H, Zhou H (2023) Detection of cigarette appearance defects based on improved YOLOv4. *Electron Res Arch* 31: 1344–1364. https://doi.org/10.3934/era.2023069
6. Wu S, Lv X, Liu Y, Jiang M, Li X, Jiang D, et al. (2024) Enhanced SSD framework for detecting defects in cigarette appearance using variational Bayesian inference under limited sample conditions. *Math Biosci Eng* 21: 3281–3303. https://doi. org/10.3934/mbe.2024145
7. Ma Y, Yuan G, Yue K, Zhou H (2023) CJS-YOLOv5n: A high-performance detection model for cigarette appearance defects. *Math Biosci Eng* 20: 17886–17904. https://doi.org/10.3934/mbe.2023795
8. Wang A, Chen H, Liu L, Chen K, Lin Z, Han J, et al. (2024) YOLOv10: Real-Time End-to-End Object Detection. *Adv Neural Inform Proc Syst* 37: 107984-108011.
9. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. (2022) LoRA: Low-Rank Adaptation of Large Language Models. *Proc ICLR* 1: 3.
10. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B (2022) High-resolution image synthesis with latent diffusion models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10674–10685. https://doi.org/10.1109/CVPR52688.2022.01042
11. Zhang K, Zhou Y, Xu X, Dai B, Pan X (2024) DiffMorpher: Unleashing the capability of diffusion models for image morphing. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7912–7921. https://doi.org/10.1109/CVPR52733.2024.00756
12. Jocher G, Chaurasia A, Qiu J (2023) Yolo by ultralytics. Available from: https://github.com/ultralytics/ultralytics
13. Li Y, Liu F (2020) Adaptive Gaussian Noise Injection Regularization for Neural Networks. In: *Proceedings of the International Symposium on Neural Networks*, 176–189. https://doi.org/10.1007/978-3-030-64221-1_16
14. Kim EK, Lee H, Kim JY, Kim S (2020) Data augmentation method by applying color perturbation of inverse PSNR and geometric transformations for object recognition based on deep learning. *Appl Sci* 10: 3755. https://doi.org/10.3390/app10113755
15. Wong SC, Gatt A, Stamatescu V, McDonnell MD (2016) Understanding Data Augmentation for Classification: When to Warp? In: *Proceedings of the International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 1–6. https://doi.org/10.1109/DICTA.2016.7797091
16. Jia S, Wang P, Jia P, Hu S (2017) Research on data augmentation for image classification based on convolution neural networks. In: *Proceedings of the Chinese Automation Congress*, 4165–4170. https://doi.org/10.1109/CAC.2017.8243510
17. Takahashi R, Matsubara T, Uehara K (2020) Data Augmentation Using Random Image Cropping and Patching for Deep CNNs. *IEEE Trans Circuits Syst Video Technol* 30: 2917–2931. https://doi.org/10.1109/TCSVT.2019.2935128

18. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. (2017) Attention is All You Need. *Adv Neural Inform Proc Syst* 30.
19. Esser P, Rombach R, Ommer B (2021) Taming transformers for high-resolution image synthesis. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 12873–12883. https://doi.org/10.1109/CVPR46437.2021.01268
20. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. (2021) Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986
21. Ba LJ, Kiros JR, Hinton GE (2016) Layer normalization. *arXiv preprint arXiv:1607.06450*.
22. Guo J, Chen X, Tang Y, Wang Y (2024) SLAB: Efficient transformers with simplified linear attention and progressive re-parameterized batch normalization. In: *International Conference on Machine Learning*, 667.
23. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. In: *International Conference on Machine Learning*, 448–456.
24. Liu S, Qi L, Qin H, Shi J, Jia J (2018) Path aggregation network for instance segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8759–8768. https://doi.org/10.1109/CVPR.2018.00913
25. Xu X, Jiang Y, Chen W, Huang Y, Zhang Y, Sun X (2022) DAMO-YOLO: a report on real-time object detection design. *arXiv preprint arXiv:2211.15444*.
26. Wang CY, Bochkovskiy A, Liao HYM (2023) YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7464–7475. https://doi.org/10.1109/CVPR52729.2023.00721
27. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet Classification with Deep Convolutional Neural Networks. *Adv Neural Inform Proc Syst* 25: 1106–1114.
28. Chen C, Zanotti Fragonara L, Tsourdos A (2020) Go wider: an efficient neural network for point cloud analysis via group convolutions. *Appl Sci* 10: 2391. https://doi.org/10.3390/app10072391
29. Zhang X, Zhou X, Lin M, Sun J (2018) ShuffleNet: An extremely efficient convolutional neural network for mobile devices. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6848–6856. https://doi.org/10.1109/CVPR.2018.00716
30. Wang C, He W, Nie Y, Guo J, Liu C, Wang Y, et al. (2023) Gold-YOLO: Efficient object detector via gather-and-distribute mechanism. *Adv Neural Inform Proc Syst* 36: 51094–51112.
31. Zhao Y, Lv W, Xu S, Wei J, Wang G, Dang Q, et al. (2024) DETRs beat YOLOs on real-time object detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16965–16974. https://doi.org/10.1109/CVPR52733.2024.01605
32. Wang CY, Yeh IH, Liao HYM (2024) YOLOv9: Learning what you want to learn using programmable gradient information. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, 1–21. https://doi.org/10.1007/978-3-031-72751-1_1
33. Khanam R, Hussain M (2024) YOLOv11: An Overview of the Key Architectural Enhancements. *arXiv preprint arXiv:2410.17725*.
34. Zhou H, Yang R, Zhang Y, Duan H, Huang Y, Hu R, et al. (2025) UniHead: Unifying Multi-Perception for Detection Heads. *IEEE Trans Neural Networks Learn Syst* 36: 9565–9576. https://doi.org/10.1109/TNNLS.2024.3412947
35. Ding X, Zhang X, Han J, Ding G (2022) Scaling Up Your Kernels to 31×31: Revisiting Large

Kernel Design in CNNs. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11953–11965. https://doi.org/10.1109/CVPR52688.2022.01166

36. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 618–626. https://doi.org/10.1109/ICCV.2017.74

AIMS Press