Pill Counting Method For Strip Plate Slots Based On YOLOv12

Qingwu Shi, Maotong Qin, Xu Du*, and Huaqi Zhao*

School of Information and Electronic Engineering, Jiamusi University, Jiamusi 154007, China

Corresponding author. E-mail: 13351649107@163.com; zhaohuaqi@126.com

Received: Aug. 14, 2025; Accepted: Sep. 30, 2025

In the modern pharmaceutical industry, automated pill counting is a critical step in the production process. However, traditional methods often fail to meet the demands of high-speed and real-time detection in terms of accuracy and efficiency. This paper, motivated by the application of strip-type pill counting machines in pharmaceutical manufacturing, proposes a pill counting and missing pill detection method within elongated strip plate holes using the YOLOv12 model. The method is capable of identifying three pill states-missing pill, single pill, and two pills vertically overlapped within a single slot-and then determines the pill quantity based on the corresponding identified states. In this study, a custom dataset was constructed, and the collected images were manually annotated. The CBAM (Convolutional Block Attention Module) attention mechanism was integrated into the YOLOv12 model to enhance its focus on small pill targets and critical regions. Additionally, negative samples were incorporated into the dataset to improve the model's ability to distinguish between background and missing pill states. With a parameter size of 2.6 M and a computational complexity of 6.4 GFLOPs, the model maintains low computational cost and lightweight characteristics while achieving high-precision detection of pill quantities and missing pill states.

Keywords: YOLOv12;Strip-type pill counting;CBAM;Object detection;Small object detection

http://dx.doi.org/10.6180/jase.202606_29(6).0010

1. Introduction

In pharmaceutical manufacturing, pills constitute one of the most prevalent dosage forms. Deficiencies such as missing pills, quantity deviations, and misaligned positioning during the packaging process not only compromise dosage accuracy and jeopardize patient safety but also cause production line disruptions and necessitate rework. These issues lead to significant economic losses and adversely impact operational efficiency as well as the enterprise's reputation for quality [1]. Traditional domestic enterprises commonly employ manual counting, mechanical counting, and photoelectric pill counting machines for pill enumeration. Manual counting is insufficient for promptly detecting and correcting missing or multiple pills in a timely manner and involves high labor intensity, often leading to worker fatigue. The most common type of mechanical counting device is a template-based mechanical pill counter, where each template is designed for a specific pill type. These templates permit only one pill per hole or groove, making it difficult to distinguish overlapping pills within a single cavity. Consequently, such devices exhibit limited adaptability and poor generalizability[2–4]. Photoelectric counting methods are widely used and rely on photoelectric sensors to detect falling pills for counting. However, these methods cannot effectively handle issues such as pill adhesion and overlap, resulting in compromised counting accuracy [5, 6].

With the advancement of visual inspection technologies, vision-based pill counting has been extensively investigated for its application in pill packaging processes[7–10]. Yao et al.[11] proposed a flat-plate pill counting machine based on machine vision, employing a target position

prediction method along detection lines. This approach is particularly suitable for single-channel, dispersed pill counting scenarios. For adhered pills, Hao et al.[12] designed a high-speed, vision-based online pill counting and packaging system. They proposed a concave point segmentation method to handle adhered pills for counting. However, this method is not applicable to scenarios involving three-dimensional overlap (e.g., vertical stacking), as the contours of the occluded lower-layer objects cannot be extracted and therefore cannot be counted. Zhang et al.[13] addressed the challenge of three-dimensional pill counting by transforming it into a two-dimensional problem. They proposed a method combining dual-threshold Otsu segmentation with Hough circle fitting to count stacked pills in top-view images, provided that the lower pills are not completely occluded by those above. Hu et al.[14] proposed a pill counting system based on an improved Faster R-CNN deep learning algorithm, achieving an accuracy of 95.47%. However, there remains room for further improvement in counting precision.

The aforementioned pill counting methods are relatively dispersed in their implementation, making them prone to issues such as friction and collision during production and packaging, which can result in pill damage. Moreover, these methods often lack effective integration with subsequent packaging processes after the pills are counted during free fall. This study proposes a method for pill counting and missing-pill detection within strip plates. The unique arrangement not only secures the position of each pill but also enables the rapid, simultaneous release of pills from a single plate into multiple bottles. When combined with a multi-channel dispensing mechanism, this approach significantly shortens the bottling cycle and enhances the overall production line throughput. Especially for the production of small pills, compared with the traditional single-channel photoelectric counting methods mentioned above, the strip plate-based counting approach can count and bottle significantly more pills within the same unit of time, offering a marked advantage in overall productivity. This results in a significantly greater overall production throughput. Compared with two-stage algorithms [15] such as Faster R-CNN [16], this study employs the YOLO (You Only Look Once) algorithm [17], which offers better detection accuracy, efficiency, and real-time performance, to detect the number of pills in strip plate holes. The YOLOv12 model is trained to achieve effective detection even for vertically overlapping pills within the holes.

2. Theory and formula

2.1. YOLOv12 Model Construction

The YOLOv12 selected in this paper is the latest generation of efficient real-time object detection model in the YOLO series. Compared with earlier versions, it introduces a deeper multi-scale feature fusion structure and a dynamic receptive field adjustment mechanism [18], enabling the model to capture more comprehensive semantic information and more distinct boundary features when addressing the pill counting task on strip plates. In addition, YOLOv12 adopts an improved RepVGGBlock structure to enhance the feature representation capability of convolutional layers, significantly improving its detection performance for small pills within the cavities. Furthermore, the improved feature fusion and decoupled structure, combined with an anchor-free detection strategy and a lightweight network architecture, reduce the computational overhead associated with anchor boxes while enhancing the model's inference speed and detection efficiency[18–20]. This effectively improves the recognition accuracy of "missing pills" and "double pills" within the strip plate holes, demonstrating strong potential for industrial applications.

2.1.1. Detection Model Based on YOLOv12

YOLOv12 offers five model variants (n, s, m, l, x) to meet different detection requirements, where " n " denotes the lightweight version and " x " represents the highest precision version. Pill detection tasks require high real-time performance, as pills move rapidly on production lines, necessitating the detection system to identify and count "missing pill(nopill), single pill(onepill), and double pills(twopills)"states within slots in an extremely short time. The lightweight version has fewer parameters and lower computational complexity, offering faster inference speeds and reduced latency. This significantly lowers computational and storage overhead while improving detection efficiency to meet real-time requirements. Furthermore, when integrating the detection model into pharmaceutical industrial control systems or embedded devices in the future, lightweight models can reduce dependency on computational resources and enhance system deployment flexibility and applicability. Therefore, this study adopts the lightweight YOLOv12n variant as the base model. Its network architecture consists of four core modules: the input layer (Input), backbone network (Backbone), neck network (Neck), and detection head network (Head).

The input layer performs preprocessing operations such as normalization and resizing to ensure input consistency and feature representation integrity. In the backbone network, YOLOv12n does not fully adopt the CSPDarkNet

from YOLOv5; instead, it is built upon an improved version of CSPDarkNet as the primary feature extraction network, combined with the innovative C 2 f module to enable cross-stage feature fusion. This module enhances feature extraction capabilities by introducing partial residual connections, strengthening the perception of targets at different scales, which is particularly effective for small pill detection on pill strips. Additionally, YOLOv12 retains the fast spatial pyramid pooling (SPPF) module to optimize and accelerate multi-scale pooling operations.

The neck network adopts a bidirectional fusion strategy based on the feature pyramid network [21] (FPN) and path aggregation network (PANet), combining top-down and bottom-up feature flows. This multi-level, multi-scale feature fusion approach preserves rich edge and texture information in images, providing better accuracy for pill detection tasks. The detection head employs a decoupled structure that separates classification from bounding box regression, reducing interference between tasks [22, 23].

In this study, the YOLOv12 model exhibited class confusion issues between the "nopill" pill state and the background, where a considerable portion of "nopill" instances were either mistaken for background or missed entirely. To address this, negative samples were added to enhance the model's discriminative ability, and a lightweight attention module, CBAM (Convolutional Block Attention Module), was integrated into the neck part of YOLOv12. While maintaining the original lightweight and fast characteristics of YOLOv12, the introduction of CBAM enhances the feature representation capabilities of the detection layers, enabling the model to automatically focus on more important image regions and channel features during training, suppress irrelevant background, and improve the discrimination between targets and background as well as the perception of pill states. The model structure is illustrated in Fig. 1.

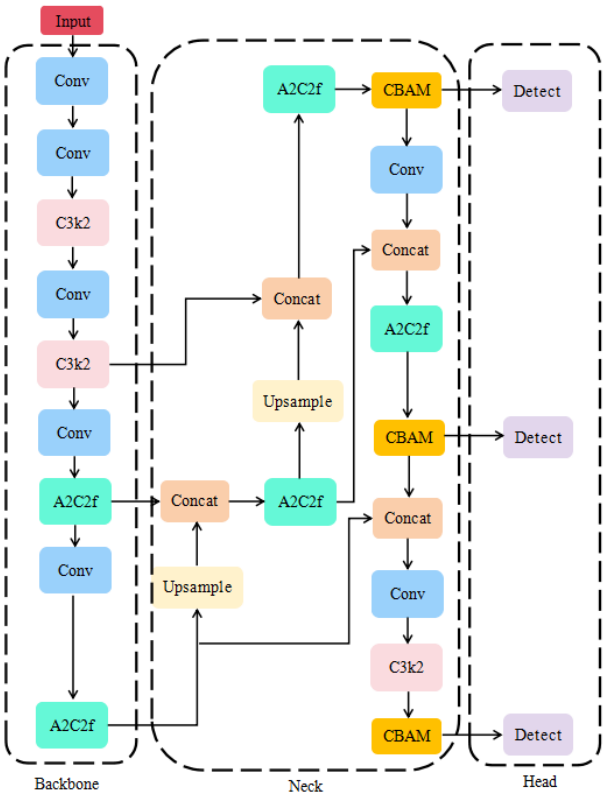


Fig. 1. YOLOv12n Network Architecture with the Integrated CBAM Module

2.2. CBAM Attention Module

The CBAM attention mechanism primarily consists of two components: the Channel Attention Module (CAM) [24] and the Spatial Attention Module (SAM). This dual attention mechanism enables the model to maintain high detection accuracy for pill locations and states under complex backgrounds while keeping the number of parameters low. The process is mathematically expressed in Eq. (1) and Eq. (2):

$$F' = M_c(F) \otimes F \tag{1}$$

$$F'' = M_s\left(F'\right) \otimes F' \tag{2}$$

F' denotes the output feature after the channel attention mechanism.

F'' represents the final refined output.

As shown in Fig. 2, CAM applies global average pooling and global max pooling to the input feature map $F \in \mathrm{R}^{C\times H\times W}$ (where H denotes the height of the feature map, W the width, and C the number of channels). Both pooled outputs are fed into a shared Multi-Layer Perceptron (MLP), and their results are then summed channel-wise. The aggregated output is passed through a Sigmoid activation to generate a one-dimensional channel attention map $Mc \in \mathrm{R}^{C\times 1\times 1}$, which enhances the response to important channels. The corresponding formula is as follows:

$$\begin{aligned} M_c(F) = \sigma\big(&\mathrm{MLP}(\mathrm{AvgPool}(F)) \\ &+ \mathrm{MLP}(\mathrm{MaxPool}(F))\big) \end{aligned} \tag{3}$$

SAM applies average pooling and max pooling along the channel dimension to the feature map refined by CAM, resulting in two spatial attention maps of size $H \times W$. These two maps are concatenated and processed with a 7×7 convolution to extract local and global spatial relationships [25, 26]. The output is then activated by a Sigmoid function to produce a two-dimensional spatial attention map $Ms \in \mathrm{R}^{l\times H\times W}$, allowing the model to more effectively

attend to salient local regions within the image [27]. The formula is as follows:

$$M_s\left(F'\right) = \sigma\left(f^{7\times7}\left(\left[\text{AvgPool}\left(F'\right);\text{MaxPool}\left(F'\right)\right]\right)\right) \quad (4)$$

The core structure of YOLOv12 detection includes three stages: feature extraction by the Backbone, feature aggregation by the Neck, and final prediction by the Head. In this study, CBAM was added to the Neck between the Backbone and Head. The dual attention mechanisms in the channel and spatial dimensions refine the Backbone's output features, and the multi-scale aggregation in the Neck further enhances the pill image features.

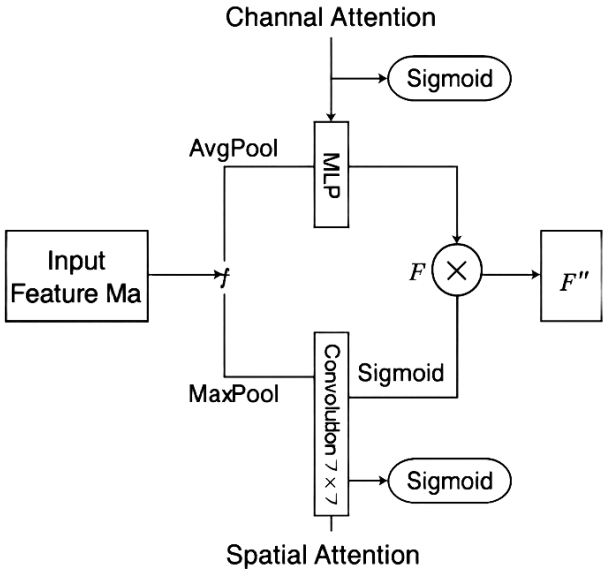


Fig. 2. Structure of CBAM Attention Mechanism

2.3. Experimental Design and Materials

2.3.1. Data Acquisition

Prior to the experiments, images of pills on strip plates were acquired under varying lighting conditions, illumination angles, and strip plate tilt angles. Subsequently, data augmentations such as rotation and blurring were applied to construct a high-quality dataset for model training and validation.

- Selection of Industrial Camera and Light Source

Industrial cameras mainly include line-scan cameras and area-scan cameras. Area-scan cameras capture images based on a pixel matrix, producing two-dimensional images [28]. In contrast, line-scan cameras capture only one line of pixels at a time, requiring continuous movement of the object or scanning by the camera to acquire multiple lines and stitch them into a complete two-dimensional image [29]. Compared with line-scan cameras, area-scan cameras offer the advantage of directly capturing two-dimensional image information, making them more widely applicable in various machine vision tasks.

Since the pills in this study are black, grayscale conversion of the original images is generally required during preprocessing. As the detection process does not rely on color features to identify the targets [30], a grayscale camera was selected for the experiments. After evaluating factors such as processing speed, cost-effectiveness, image quality, and the specific detection requirements, a monochrome area-scan industrial camera was chosen. This type of camera captures high-quality grayscale images, making it well-suited for the image acquisition tasks required in this work.

Lighting is equally crucial for image acquisition. For the strip plate pills targeted in this study, bar-shaped LED lighting offers significant advantages over other light sources such as halogen and fluorescent lamps, including lower energy consumption, longer service life, and more stable light output. As a result, it has become the mainstream choice in machine vision inspection and industrial illumination [31]. Therefore, this study adopts a bar-shaped LED light source.

3. Negative sample introduction

Negative samples refer to images that do not contain any of the target objects to be identified in the task. In this study, negative samples are defined as images that do not include any pill states (nopill, onepill, or twopills) and have no annotated bounding boxes. Their corresponding label files are empty and contain no annotation information. Examples of negative sample images are shown in Fig. 3 and include the following scenarios: (1) Flat plate regions without any holes or pills, including the edges of the strip plate. (2) Pure background regions, such as the floor. (3) Other irrelevant objects, such as strip-shaped metal components or smooth surfaces.

Negative samples enable the model to learn the distribution characteristics of images without target objects, thereby improving its ability to distinguish between background and targets (especially the "nopill" state) and reducing false-positive detections.

4. Image acquisition

During image acquisition, the camera was mounted on a fixed stand approximately 30 cm above the strip plate to ensure stable and repeatable image data. The target was

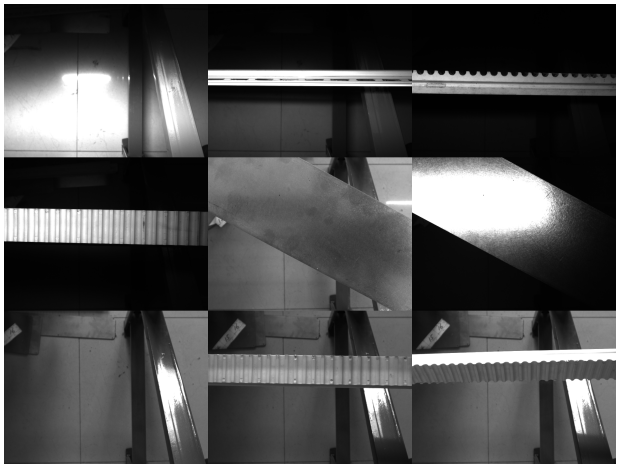


Fig. 3. Examples of negative sample images

a long strip plate containing two rows of holes, with each hole potentially containing one of three pill states: missing pill(nopill), single pill(onepill), or double pills(twopills). Considering the complex and variable lighting conditions in actual production environments, where illumination angles and intensities significantly affect image quality, this study enhanced dataset diversity and model generalization by acquiring images under both artificial lighting and natural light. Artificial lighting conditions included vertical illumination, high-angle side illumination (with the light source positioned at a $60-90^\circ$ angle above the object), and low-angle side illumination (forming a $0-45^\circ$ angle with the object). Additionally, under each lighting condition and natural light, images were captured with the strip plate tilted at various angles, including $10^\circ, 20^\circ, 30^\circ$, and 40°, to simulate perspective variations caused by conveyor vibrations or changes in plate placement angles in production lines. The image acquisition hardware setup is shown in Fig. 4, and example images under different acquisition conditions are shown in Fig. 5.

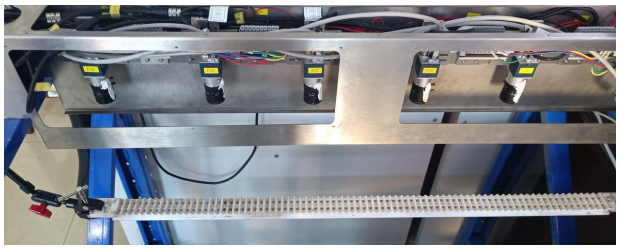


Fig. 4. Image Acquisition Environment

4.1. Dataset Construction

A portion of the collected images underwent data augmentation operations, including rotation, blurring, exposure

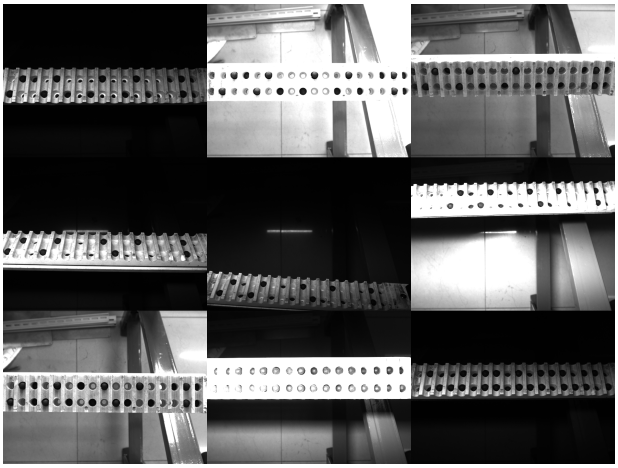


Fig. 5. Image Acquisition Results of Pills on Different Strips

level adjustment, and the mosaic augmentation method supported by YOLOv12. Negative samples were also incorporated to improve the model's ability to distinguish non-target scenes. In total, 1,481 images were generated. All images were manually annotated using LabelImg, with bounding boxes tightly enclosing the pill areas within the holes to ensure clear boundaries and accurate class labels. The annotation results were saved in YOLO format as text files, with each image corresponding to a single label file containing the class number and the normalized position of each target within the image. After annotation, the dataset was split into training, validation, and test sets in a $7:2:1$ ratio to ensure a balanced distribution of samples during model training. The annotation results are shown in Fig. 6.

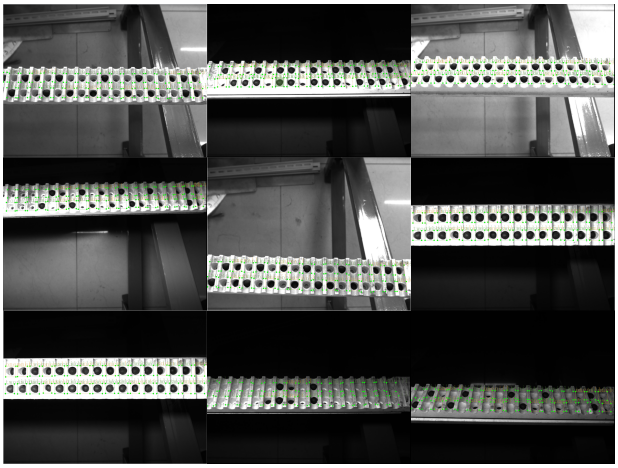


Fig. 6. Annotation Results of Strip Plate Pills

5. Experimental setup

5.1. Experimental Environment

The experiments were conducted on a Windows 10 (64) operating system using PyTorch 2.6.0 as the deep learning framework, with CUDA 12.4 installed for acceleration. Model training was implemented based on Ultralytics YOLOv12. The computing platform was a laptop equipped with an NVIDIA GeForce MX350 GPU with 2 GB of video memory, an Intel Core i5 CPU, and 16 GB of system memory (RAM). The main training parameter settings are shown in Table 1.

5.2. Evaluation Metrics

The primary evaluation metrics used in this study are as follows:
(1) Precision (P): The proportion of correctly detected targets among all predicted targets.

$$P = \frac{TP}{TP + FP} \tag{5}$$

TP - The number of pills correctly identified as their corresponding class.
FP - The number of samples predicted as the current class but actually belonging to other classes.
(2) Recall (R): The proportion of all true targets that are correctly detected.

$$R = \frac{TP}{TP + FN} \tag{6}$$

FN - The number of samples that truly belong to the current class but are predicted as other classes.
(3) F1 Score: The harmonic mean of precision and recall.

$$F_1 - \text{ Score } = 2 \times \frac{P \times R}{P + R} \tag{7}$$

(4) mAP@0.5: Measures the overall performance of the model by combining precision and recall.

$$AP = \int_0^l P(R)\mathrm{d}R \tag{8}$$

$$mAP@0.5 = \frac{\sum_{i=1}^{C} AP_i}{C} \tag{9}$$

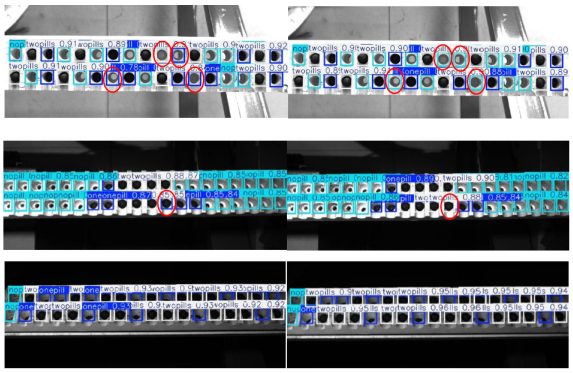
C - The total number of pill classes.

APi -The average precision for class i, equal to the area under the precision-recall curve for that class, where i denotes the number of detections.
(5) Number of Parameters: The total number of trainable parameters in the model.
(6) GFLOPs: The number of floating-point operations, reflecting model complexity.

6. Result and discussions

After completing the training, this section compares the detection results of the model before and after the introduction of the CBAM attention mechanism, analyzes various evaluation metrics including the changes in false positives (FP) and false negatives (FN) of the YOLOv12n-CBAM model with and without negative samples, and conducts comparative experiments with other models on the same dataset - with the goal of comprehensively evaluating the performance of the strip plate pill detection model.

6.1. Model Detection Results Comparison



(a) Detection Results of YOLOv12n Model (b) Detection Results of YOLOv12n Model with CBAM

Fig. 7. Comparison of Detection Results Between the Two Models

Multiple sets of detection experiments were conducted using the YOLOv12n model and the CBAM-enhanced model. The comparison of detection results is shown in Fig. 7. It can be observed that YOLOv12n exhibited certain false detection phenomena, specifically: misclassifying missing pill areas as containing single or double pills, and simultaneously detecting both "nopill" and "twopills" bounding boxes within a single empty hole. Such cases prevent accurate recording of the number of pills within each hole. In contrast, after adding the CBAM module, the detection results showed that all pill targets were accurately identified without any false detections or missed detections. The model correctly detected the pill quantity states, and its performance remained excellent on previously unseen data, demonstrating good generalization capability.

6.2. Model Training Results

The variations of evaluation metrics during the training process are illustrated in Fig. 8, and the average values of key indicators are presented in Table 2. In the early stages of training, the model primarily focused on adjusting the predicted bounding boxes. As the number of training epochs

Table 1. Model Training Parameter Settings

Parameter	Value	Parameter	Value
epochs	300	lr0	0.01
bacth	4	optimizer	SGD
imgsz	640	workers	0
Close_mosaic	10	weight_decay	0.0005

Table 2. Precision, Recall, F1Score, mAP@0.5, and FPS

Evaluation Metric	Precision/%	Recall/%	$F1_{score}$ /%	mAP@0.5/%	FPS/ (FPS/s)
Value	99.21	99.26	99.2	99.28	36

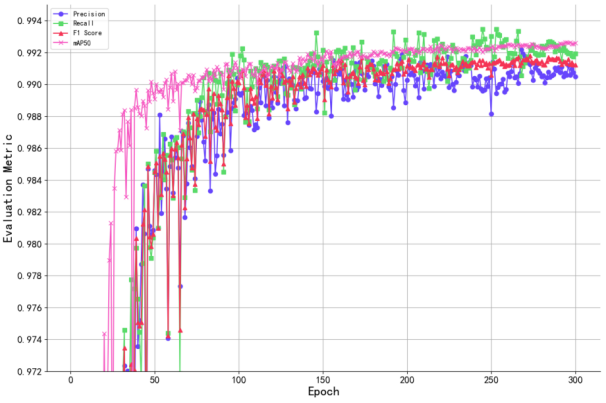


Fig. 8. Variation of Model Evaluation Metrics

increased, the precision and recall gradually stabilized at approximately 0.9921 and 0.9926 , respectively. This indicates that the model achieved a prediction accuracy of 99.21% for detected pill states, demonstrating strong precision. Meanwhile, 99.26% of the actual targets were correctly identified, suggesting a low rate of missed detections. Furthermore, the model's performance metrics exhibited minimal fluctuation during convergence, indicating strong training stability and consistent convergence behavior, and ensuring high reliability of the prediction results.

The model's F1-score approaches 0.992 , which is the harmonic mean of precision and recall. Values closer to 1 indicate a stronger balance between precision and recall. The current score demonstrates that the model achieves both high accuracy and high coverage in the detection of all pill states, reflecting its balanced performance and robust overall detection capability.

The test results show that the model achieves an inference speed of 36 FPS, processing approximately 36 image frames per second. This frame rate helps reduce detection latency during the high-speed operation of pill strip plates and enhances system responsiveness, demonstrating a certain degree of real-time detection capability. In terms of deployment adaptability, the model is capable of meeting future expansion needs on host computers or edge computing devices, making it suitable for automated application scenarios. mAP@0.5 refers to the mean average precision calculated at an IoU threshold of 0.5 , used to evaluate whether the model can correctly identify targets in detection tasks. Only when the overlap between the predicted box and the ground truth box exceeds 50% is it considered a correct detection [32]. The model achieves a mAP@ 0.5 of 0.9928 , surpassing the industrial benchmark requirement of 0.95 . This indicates that the model has strong capability in basic detection tasks, being able to stably identify targets and output detection results with high confidence.

The F1-confidence curves are shown in Fig. 9. Within the high-confidence interval of 0.1 to 0.8 , the F1 scores for each category- "onepill", "nopill", "twopills"-as well as the overall (All) score, remain at a high level. This indicates that the model has successfully learned discriminative features between targets and backgrounds during training, achieving a good balance between precision and recall, and can stably output high-quality predictions, demonstrating effective learning capability. For all three pill quantity categories, the model exhibited highly consistent performance. Such balance is crucial in practical industrial applications, ensuring that the system not only accurately identifies target pills but also effectively avoids misdetections that could lead to counting errors. Overall, the model performed excellently at non-extremely high confidence levels, and in practical applications, the confidence threshold should be set appropriately by balancing confidence with prediction performance.

FP (False Positive): The number of samples predicted as the current class but actually belonging to other classes.

FN (False Negative): The number of samples that truly belong to the current class but are predicted as other classes.

As shown in Table 3, after incorporating negative samples, the model exhibited a reduction in both false positives (FP) and false negatives (FN) across all pill states. Specif-

Table 3. Comparative Results Before and After Negative Samples Integration

Pill states	Metric	Before	After	Change
onepill	FP	18	13	-5
	FN	20	16	-4
nopill	FP	93	65	-28
	FN	16	10	-6
twopills	FP	16	15	-1
	FN	15	14	-1

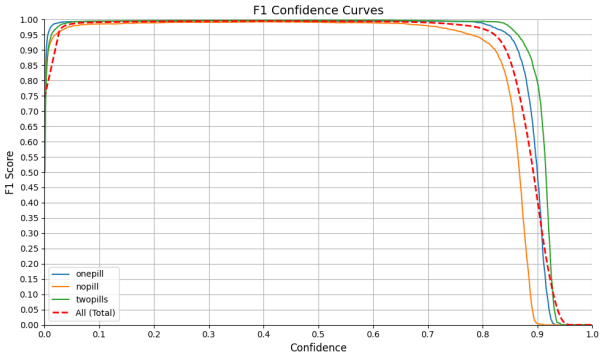


Fig. 9. F1 Score-Confidence Curve

ically, the total FP decreased by 34 and the total FN decreased by 11, indicating an overall improvement in performance and more stable predictions across different sample types. Notably, the largest decrease was observed in the "nopill" category, demonstrating that the inclusion of negative samples effectively reduced the confusion between the missing pill state, background, and other classes.

6.3. Comparative Experiments with Different Models

Based on the comparison results in Table 4, the introduction of the CBAM module into YOLOv12n led to slight improvements in model performance, with precision, recall, and mAP@0.5 increasing by 0.49%, 0.37%, and 0.27%, respectively. Compared with other models, YOLOv12n-CBAM achieved outstanding performance in all evaluation metrics. Although its parameter size and GFLOPs were not the smallest- 6.4 M and 2.6 G respectively-the increase in parameter size compared to YOLOv12n was minimal, and the increase in floating-point operations remained within a reasonable range. These results indicate that the CBAM module effectively enhanced the model's target detection capability without significantly increasing computational burden or complexity, achieving an optimal balance between performance and complexity.

7. Conclusion

To meet the demand for automated missing pill detection and pill counting in pharmaceutical strip plates during the packaging process, this study achieved the following:

- Dataset Construction

Established a representative pill image dataset tailored for strip plate inspection tasks, ensuring diversity in pill states and backgrounds.

- Model Development

Based on the YOLOv12n architecture, a lightweight object detection model was designed and trained by integrating the CBAM attention mechanism, incorporating negative samples, and applying a variety of image augmentation strategies.

- Performance Achievements

Experimental results showed that the model achieved a detection accuracy of 99.21%, accurately identifying pill quantities within strip plate holes, with excellent performance in recall(0.9926), F1-score(0.992), and mAP@0.5(0.9928). After introducing negative samples, the model exhibited a notable reduction in both false positives (FP) and false negatives (FN) for the detection of pill states, decreasing by 34 and 11 cases respectively. This enhancement further improved the model's ability to distinguish between pill states and background regions. The model has 2.6 M parameters and 6.4 GFLOPs , achieving an inference speed of 36 FPS , with notable advantages in terms of lightweight design, computational efficiency, and real-time performance. Training and inference could be performed on mid-to-low-end GPUs (NVIDIA MX350), indicating potential for edge deployment and suitability for industrial production line real-time detection and host computer deployment requirements.

- Future Work

However, the actual scenarios of pharmaceutical production inspection are highly complex and diverse, with pills varying widely in type and arrangement patterns. The current model still requires further optimization to improve its generalization ability and adaptability. Future work will

Table 4. Experimental Results of Different Models on This Dataset

Evaluation Metric	Precision/%	Recall/%	GFLOPs/G	Model Size/M	mAP@0.5/%
SSD	96.02	14.46	60.95	23.84	86.87
Faster-RCNN	74.56	91.59	370.2	137.1	90.29
YOLOv5n	98.91	97.62	2.5	7.1	97.83
YOLOv8n	97.24	97.87	3.01	8.2	98.02
YOLOv10n	95.66	96.29	2.7	8.2	96.93
YOLOv12n	98.72	98.89	2.5	5.8	99.01
YOLOv12n-CBAM	99.21	99.26	2.6	6.4	99.28

focus on developing multi-task integrated intelligent detection models for the entire pharmaceutical production process, achieving comprehensive identification and statistical analysis of appearance quality and defect types.

8. Acknowledgements

This work was supported by Key R&D Program (Innovation Base) Project of Heilongjiang Province (Project No. JD24A014), 2024 Heilongjiang Province Higher Education Teaching Reform Research General Project (Project No. SJGYY2024227), Heilongjiang Provincial Department of Education Innovation Team Project (Project No. 2024-KYYWF-0625), and Jiamusi University "Dongji" Academic Team Project (Project No. DJXSTD202417).

References

[1] X. Yao, (2021) *"Current Situation and Development Prospects of Traditional Chinese Medicine Industry"* **Chinese Traditional and Herbal Drugs** ***52***(17): 5115–5119. DOI: CNKI:SUN:ZCYO.0.2021-17-001.

[2] Z. Fan, (2022) *"Design of an Automatic Pill Counting System Based on a Single-Chip Microcomputer"* **Plant Maintenance Engineering** (09): 113–114. DOI: 10.16621/j.cnki.issn1001-0599.2022.05.49.

[3] Y. Ruidong. *"Research on High-Precision Pill Counting Machine Control System Based on ARM"*. (mathesis). Shandong University, 2015.

[4] J. Yang, C. Dou, L. Xin, W. Liu, and X. Zhou, (2018) *"Research on tablet granule counting algorithm based on visual matching technology"* **Packaging Engineering** ***39***(19): 175–180. DOI: 10.19554/j.cnki.1001-3563.2018.19.031.

[5] G. Liu. *"Design of Counting Machine Control System Based on DSP and PLC"*. (mathesis). Nanchang Hangkong University, 2013.

[6] Z. Wang, J. Chen, and R. Ai, (2017) *"Development Trend and Application Prospect of Photoelectric Detection Technology"* **Sichuan Cement** (03): 152. DOI: CNKI:SUN:SCSA.0.2017-03-149.

[7] Q. Sun and J. Cai, (2020) *"Detection System of Flat Plate Counting Machine Based on FPGA"* **Light Industry Machinery** ***38***(03): 69–73. DOI: CNKI:SUN:QGJX.0.2020-03-014.

[8] C. Phromlikhit, F. Cheevasuvit, and S. Yimman. "Tablet counting machine base on image processing". In: *The 5th 2012 Biomedical Engineering International Conference*. 2012, 1–5. DOI: 10.1109/BMEiCon.2012.6465508.

[9] J. Moon, S. Lim, H. Lee, S. Yu, and K.-B. Lee, (2022) *"Smart Count System Based on Object Detection Using Deep Learning"* **Remote Sensing** ***14***(15): 3761. DOI: 10.3390/rs14153761.

[10] A. D. Nguyen, H. H. Pham, H. T. Trung, Q. V. H. Nguyen, T. N. Truong, and P. L. Nguyen, (2023) *"High accurate and explainable multi-pill detection framework with graph neural network-assisted multimodal data fusion"* **Plos One** ***18***(9): e0291865. DOI: 10.48550/arXiv.2303.09782.

[11] Y. Yao, J. Cai, and Q. Liu, (2018) *"Detection method of flat plate counting machine based on machine vision"* **Optical Instruments** ***40***(04): 9–14. DOI: CNKI:SUN:GXYQ.0.2018-04-002.

[12] M. Hao, K. Sun, T. Liu, and G. Wang, (2023) *"Design of high-speed online pill counting and packaging system based on machine vision"* **Techniques of Automation and Applications** ***42***(06): 38–40. DOI: 10.20033/j.1003-7241.(2023)06-0038-03.

[13] J. Zhang and W. Zhu, (2018) *"Counting of circular overlapping particles based on depth image processing technology"* **Information Technology** (06): 71–75+80. DOI: 10.13274/j.cnki.hdzj.2018.06.015.

[14] A. Hu and Z. Li, (2018) *"Counting machine system based on improved Faster R-CNN"* **Packaging Engineering** ***39***(09): 141–145. DOI: 10.19554/j.cnki.1001-3563.2018.09.025.

[15] H.-J. Kwon, H.-G. Kim, and S.-H. Lee, (2022) *"Pill Detection Model for Medicine Inspection Based on Deep Learning"* **Chemosensors *10***(1): DOI: 10 . 3390 / chemosensors10010004.

[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik. "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014, 580–587. DOI: 10.1109/CVPR.2014.81.

[17] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. "You Only Look Once: Unified, Real-Time Object Detection". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, 779–788. DOI: 10.1109/CVPR.2016.91.

[18] J. Ma, Y. Zhou, Z. Zhou, Y. Zhang, and L. He, (2025) *"Toward smart ocean monitoring: Real-time detection of marine litter using YOLOv12 in support of pollution mitigation"* **Marine Pollution Bulletin *217***: 118136. DOI: https://doi.org/10.1016/j.marpolbul.2025.118136.

[19] J. Bu. *"Research on PCB surface defect detection method based on improved YOLOX"*. (mathesis). Liaoning University of Science and Technology, 2023. DOI: 10 . 26923/d.cnki.gasgc.2023.000088.

[20] R. Khanam and M. Hussain, (2025) *"A Review of YOLOv12: Attention-Based Enhancements vs. Previous Versions"* **arXiv *arXiv:2504.11995***: DOI: https://doi.org/10.48550/arXiv.2504.11995.

[21] H. W. Ting, S. L. Chung, C. F. Chen, et al., (2020) *"A Drug Identification Model Developed Using Deep Learning Technologies: Experience of a Medical Center in Taiwan"* **BMC Health Services Research *20***: 312. DOI: 10.1186/s12913-020-05166-w.

[22] J. Chen and X. Wang, (2024) *"Dense small object detection algorithm for UAV aerial images based on improved YOLOv5"* **Computer Engineering and Applications *60***(03): 100–108.

[23] R. Sapkota, M. Flores-Calero, R. Qureshi, et al., (2025) *"YOLO advances to its genesis: a decadal and comprehensive review of the You Only Look Once (YOLO) series"* **Artificial Intelligence Review *58***: 274. DOI: 10.1007/s10462-025-11253-3.

[24] S. Yin, L. Wang, M. Shafiq, L. Teng, A. A. Laghari, and M. F. Khan, (2023) *"G2Grad-CAMRL: An Object Detection and Interpretation Model Based on Gradient-Weighted Class Activation Mapping and Reinforcement Learning in Remote Sensing Images"* **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing *16***: 3583–3598. DOI: 10.1109/JSTARS.2023.3241405.

[25] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon. "CBAM: Convolutional Block Attention Module". In: *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part VII*. 2018, 3–19. DOI: 10.1007/978-3-030-01234-2_1.

[26] Q. Shi, S. Yin, K. Wang, et al., (2022) *"Multichannel convolutional neural network-based fuzzy active contour model for medical image segmentation"* **Evolving Systems *13***: 535–549. DOI: 10.1007/s12530-021-09392-3.

[27] C. Zhang. *"Research on rotated object detection method based on convolutional neural networks"*. (mathesis). University of Electronic Science and Technology of China, 2023. DOI: 10 . 27005 / d . cnki . gdzku . 2023 . 002432.

[28] Y. Pan. *"Research on FPGA acceleration technology of feature extraction in visual inspection"*. (phdthesis). Hefei University of Technology, 2021. DOI: 10.27101/d.cnki.ghfgu.2021.000005.

[29] X. Liang. *"Research on pill coating defect detection technology based on machine vision"*. (mathesis). Chongqing University of Science and Technology, 2023. DOI: 10.27854/d.cnki.gcqkj.2023.000357.

[30] Z. Wu. *"Research and application of online pill defect detection system based on machine vision"*. (mathesis). Tianjin Polytechnic University, 2023. DOI: 10.27357/d.cnki.gtgyu.2023.001211.

[31] L. Liang. *"Research on particle counting and defect detection system based on visual tracking"*. (mathesis). South China University of Technology, 2014.

[32] W. Zhou, B. Sun, L. Shi, and S. Yang, (2025) *"Research on potato leaf disease detection method based on YOLO model"* **Automation Instrumentation *40***(05): 71–75. DOI: 10.19557/j.cnki.1001-9944.2025.05.014.