Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Random Rolling Attention Augmentation for Efficient Agricultural Disease and Pest Detection

Weijuan Han
Zhongyuan Institute of Science and Technology
Yuwen Ding
Henan Medical College
Jinglong Kang
Henan Provincial Public Security Department
Aojie Zhao
Shaanxi University of Technology
Xinjie Dong
dong3841919@163.com

Henan Provincial Public Security Department

Research Article

Keywords: agricultural disease and pest detection, YOLOv12, random rolling transformer, attention mechanism, mAP, GFLOPS

Posted Date: April 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9318824/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Corresponding author: Xinjie Dong

Random Rolling Attention Augmentation for Efficient Agricultural Disease and Pest Detection

Weijuan Han[1], Yuwen Ding[2], Jinglong Kang[3], Aojie Zhao[4], Xinjie Dong[*5]

[1]School of Mechanical and Electronic Engineering, Zhongyuan Institute of Science and Technology, No. 50 Zhengkai Avenue, Zhengzhou, 450000, Henan, China.
[2]Editorial Department, Henan Medical College, No. 8 Shuanghu Avenue, Zhengzhou, 451191, Henan, China.
[3]Technical Investigation Division, Henan Provincial Public Security Department, Qiancheng Avenue, Zhengzhou, 450046, Henan, China.
[4]TrineEngineering Institute, Shaanxi University of Technology, No. 1 East 1st Ring Road, Hanzhong, 723001, Shaanxi, China.
[5*]Science and Technology Information Division, Henan Provincial Public Security Department, Qiancheng Avenue, Zhengzhou, 450046, Henan, China.

Contributing authors: hanweijuan@zykj.edu.cn; 545734761@qq.com; abner.kang.dev@gmail.com; 3481409647@qq.com; dong3841919@163.com;

Abstract

Accurate detection of agricultural diseases and pests is essential for crop protection and food security worldwide. Real-world field applications face challenges including small objects, complex backgrounds, high class similarity, and limited computational resources. This work proposes a Random Rolling Transformer (RRT), which introduces random circular shifts along channel and sequence dimensions into multi-head self-attention to enrich feature interactions without increasing parameters or computation. Integrated into YOLOv12, the proposed RRT-YOLO is evaluated on the IP102 pest dataset and a tomato leaf disease dataset. Results show that RRT-YOLO improves mAP@50 by 2.5% and mAP@50–95 by 3.9% on IP102, and by 8.8% and 4.2% on the tomato disease dataset, while maintaining identical model size and complexity. This attention perturbation strategy offers an effective and efficient solution for lightweight agricultural vision detection and can be extended to other visual computing tasks. The code and detailed descriptions can be accessed via the following repository: https://github.com/glorioustory/Random-Rolling-Attention-Augmentation.git.

Keywords: agricultural disease and pest detection, YOLOv12, random rolling transformer, attention mechanism, mAP, GFLOPS

1 Introduction

Agricultural diseases and pests pose a persistent threat to global food security, causing substantial yield losses in staple crops such as rice, wheat, and maize. In major grain-producing regions such as Henan Province, China, timely and accurate detection is essential for effective management and prevention. Traditional monitoring methods, typically based on manual inspection, are labor-intensive, time-consuming, and difficult to scale to large fields or real-time scenarios.

Recent advances in computer vision and deep learning have enabled automated disease and pest detection systems, providing scalable, real-time, and reliable monitoring [1, 2]. Among these approaches, one-stage detectors (e.g., the YOLO series) are popular due to their favorable trade-off between speed and accuracy, which facilitates deployment on embedded devices, drones, and mobile platforms.

PestLite [3] modified YOLOv5 by designing a Multi-Level Spatial Pyramid Pooling Fusion module. This method improved mAP@50 from 87.9% to 90.7% while reducing parameters and enabling real-time pest detection. YOLO-Pest [4] replaced standard backbone modules with a CAC3 module. The model achieved 91.9% mAP@0.5 on the Teddy Cup pest dataset, outperforming YOLOv5s by nearly 8% in mAP@50. For dense small objects in remote sensing, a denoising FPN with Trans R-CNN [5] was proposed, where contrastive learning suppressed noise across fused feature levels before Transformer-based detection, significantly improving tiny-object detection.

MS-P2-YOLO [6], an improved YOLOv8s-based small-target detection algorithm, was evaluated on VisDrone2019. The results showed that precision, recall, and mAP@50 increased by 9.0%, 12.9%, and 12.9%, respectively, compared with YOLOv8s, while the number of parameters decreased by 4%. A ROI-aware multiscale cross-attention Vision Transformer [7] was proposed for pest image classification and detection on IP102, using a dual-branch design and multiscale cross-attention fusion. MSFNet-CPD [8] introduced a multiscale cross-modal fusion network for crop pest detection. Experiments demonstrated that MSFNet-CPD consistently outperformed state-of-the-art methods on multiple pest detection benchmarks. Xiong [9] utilized background subtraction to provide YOLO with the location and features of small dynamic targets, thereby reducing the missed detection rate for small targets. The proposed method improved the precision by 2.3%, recall by 3.5%, and F1-score by 3.1%.

Recent studies have further explored improved detection frameworks for agricultural diseases and pests. For rice pest detection, lightweight and improved YOLOv8 variants have been proposed to enhance accuracy while maintaining real-time performance [10], and enhanced YOLOv10-based frameworks have been developed for efficient pest monitoring in smart agriculture [11]. In addition, transformer-based end-to-end detectors have also been applied to rice pest detection, demonstrating the feasibility of attention-centric architectures in this domain [12]. For broader crop scenarios, recent work has reported improved YOLO-based models for joint corn leaf disease and pest detection [13].

Transformer-based architectures [14] have shown promise in object detection, particularly for capturing long-range dependencies and global context, which can help distinguish visually similar categories and detect small, low-contrast objects [15]. Nevertheless, mature Transformer detectors often require substantial computation and many training epochs to converge, and their accuracy may degrade in complex contexts, limiting their practicality for agricultural disease and pest detection.

Many recent object detection studies [16] focus on convolutional modifications (e.g., improved pooling, attention, and feature fusion). Although Transformer-based methods can improve performance, they often incur additional computational cost or require heavier models. Our Random Rolling Transformer (RRT) addresses this gap by introducing randomized circular shifts to the query (Q), key (K), and value (V) tensors along the channel and sequence dimensions. RRT acts as a form of attention augmentation, encouraging the model to learn more robust and diverse representations. Our contributions are summarized as follows:

- We propose a Random Rolling Transformer (RRT), an attention perturbation mechanism that introduces channel- and sequence-wise random shifts to improve detection accuracy without increasing parameters or FLOPs.
- We integrate RRT into YOLOv12 to build a more effective detector (RRT-YOLO). On a curated subset of IP102, RRT-YOLO improves mAP@0.5 by 2.5% and mAP@0.5–0.95 by 3.9% over the baseline, with no additional computational burden. On the tomato leaf disease dataset, RRT-YOLO improves mAP@50 by 8.8% and mAP@50–95 by 4.2% respectively.

The remainder of this paper is organized as follows: Section 2 describes YOLOv12 and the proposed method; Section 3 presents the datasets and experimental results; Section 4 discusses comparisons with other methods; and Section 5 concludes the paper and outlines future work.

2 Materials and methods

2.1 Regional Overview

Henan Province is situated in central China (31–36°N, 110–116°E) and spans the middle and lower reaches of the Yellow River. Henan has long been among Chinese top grain-producing provinces, with wheat, maize, and rice as major crops (Fig. 1). According to the National Bureau of Statistics, Henan's total grain output reached 135 billion jin (67.5 million tons) in 2025, ranking second nationwide. The total grain output has remained above 130 billion jin (65 million tons) for nine consecutive years.

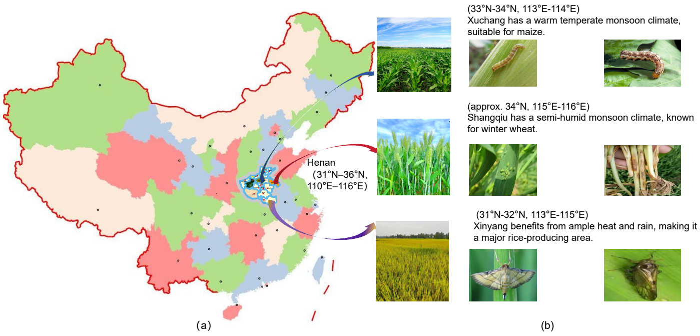

Fig. 1 Henan Province. (a) Location of Henan in China; (b) main producing areas of three major grain crops in Henan.

2.2 YOLOv12n

YOLOv12 is a breakthrough in real-time object detection, featuring an attention-centric architecture [17, 18]. As one of the latest YOLO (You Only Look Once) models, YOLOv12 integrates attention mechanisms into its core design to better balance computational efficiency and detection accuracy in real-time applications. As shown in Fig. 2, the YOLOv12 architecture comprises three primary components: (i) a backbone network for feature extraction, (ii) a neck network for multi-scale feature fusion, and (iii) a head network for final detection output. The backbone incorporates advanced modules such as A2C2f (Area Attention-enhanced Cross-stage Partial Network with Two Fusions), which integrates the proposed Area Attention (A2) mechanism.

YOLOv12 introduces several key innovations. First, it introduces the Area Attention (A2) mechanism. By dividing feature maps into four or more equal regions, the model reduces attention complexity while retaining global context awareness. Second, the Residual Efficient Layer Aggregation Network (R-ELAN) enhances feature aggregation through block-level residual connections and scaling strategies. Third, FlashAttention is integrated to optimize memory access patterns and accelerate attention computations on modern GPU architectures [19].

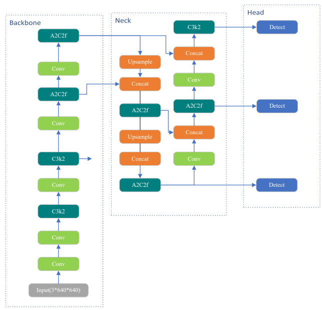

Fig. 2 Structure of YOLOv12.

2.3 Random Rolling Transformer

Area Attention in YOLOv12 is not a completely new attention formula; instead, it is an architectural strategy based on standard scaled dot-product attention. It improves efficiency and captures information at different scales via patch partitioning and intra-/inter-region attention fusion. The computational complexity of standard self-attention is quadratic in the sequence length, i.e., $O(N^2)$; when the image resolution is high, N can be very large. The core idea of Area Attention is to divide the feature map into several non-overlapping regions and compute attention independently within each region, thereby reducing the overall computational cost. The formulation is given in Eq.(1)–Eq.(3), where F_{in} denotes the input feature map, P_i denotes a partition (region) of the feature map, Q, K, and V are the query, key, and value tensors, D_k is the feature dimension of key in each head, H denotes the dimension of heads, X_i is the output feature corresponding to P, and F_{out} denotes the output feature map.

$$F_{\text{in}} = \{P_1, P_2, \ldots, P_n\}, \quad P_i \in \mathbb{R}^{B\times C\times H_i\times W_i} \tag{1}$$

$$X_i = V \times \left(\text{Softmax}\left(\frac{Q^T K}{\sqrt{D_k}}\right)\right)^T, \quad Q, K, V \in \mathbb{R}^{B\times H\times D_k\times N} \tag{2}$$

$$F_{\text{out}} = \{X_1, X_2, \ldots, X_n\} \tag{3}$$

In the proposed method named Random Rolling Transformer, we introduce a random circular shift augmentation to the Q/K/V tensors. As shown in Eq.(4)–Eq.(7), ΔN denotes the rolling value of the sequence dimension, ΔD_k denotes the rolling value of the head dimension, S_i represents the range of variation, T_i represents a batch of feature maps to be calculated, B represents the batch size, H represents the number of heads, and R represents the computed and integrated feature map. For each batch, we randomly shift parameters ΔN and ΔD_k from a predefined range (e.g., $[-3, 3]$). A two-dimensional circular shift is then applied to the N and D_k of the tensor using modular arithmetic, effectively augmenting the data representation within the attention mechanism.

$$\Delta N = \text{RandomChoice}(S_1), \quad S_1 \in (-n, n),\ n \in \mathbb{Z} \tag{4}$$

$$\Delta D_k = \text{RandomChoice}(S_2), \quad S_2 \in (-m, m),\ m \in \mathbb{Z} \tag{5}$$

$$T_i = T_i\Big(B, H, (D_k + \Delta D_k) \bmod D_k, (N + \Delta N) \bmod N\Big) \tag{6}$$

$$R = [T_1; T_2; \cdots; T_B], \ R \in \mathbb{R}^{B\times H\times D_k\times N} \tag{7}$$

In each training batch, the spatial sequence dimension N is randomly rolled, allowing each token to interact with its local neighborhood more effectively. This increases the effective receptive field and helps the model capture richer contextual information. Meanwhile, random rolling along the channel dimension D_k promotes the mixing of representations learned for the target task, thereby improving capacity of feature extraction. A schematic illustration of rolling along the channel and sequence dimensions is shown in Fig. 3.

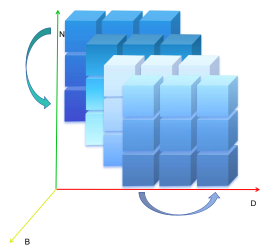

Fig. 3 Schematic diagram of the working principle of Random Rolling Transformer. D represents the channel dimension, N represents the sequence dimension, and B represents the batch.

3 Results

3.1 Experimental Setup

The experimental setup included a 13th Gen Intel(R) Core(TM) i5-13600KF (3.5 GHz, 14 cores), 32 GB of RAM, and an RTX 4060 Ti GPU with 16 GB of VRAM. The software environment included Windows 11, Python 3.11, Torch 2.6.0, CUDA 12.4, and PyCharm 2023.3. Each model was trained for 150 epochs with a batch size of 16. We used SGD as the optimizer, with an initial learning rate of 0.01, momentum of 0.937, and weight decay of 0.0005. Other parameters included close_mosaic set to 0 and mosaic set to 1.0. FlashAttention was not enabled during training. The weights achieving the highest mAP on the validation set were selected for evaluation.

3.2 Pest Dataset

The IP102 dataset is a comprehensive image collection of agricultural pests, originally containing 102 insect species [20]. We use a publicly available YOLO-format release of IP102 for training and evaluation [21]. Due to its focus on the agricultural region of Henan Province, a major production area for wheat, maize, and rice, this study exclusively utilizes the subset of 36 pest species that specifically infest these three crops. The selected pests include key economically significant species, which are prevalent in Henan's cropping systems. This subset ensures relevance to regional agricultural practices while maintaining ecological diversity within the dataset. The training set contains 5966 images, and the validation set contains 453 images.

3.3 Selection of Rotation Parameters

As shown in Table 1, D and N denote the random rotation magnitudes along the channel and sequence dimensions, respectively. A series of ablation experiments were conducted to investigate the impact of varying the values of D and N on detection performance, while keeping the same parameters and computational complexity. Detection performance was quantified using two standard metrics: mAP@50 (mean Average Precision at an IoU threshold of 0.5) and mAP@50–95 (mean Average Precision across IoU thresholds from 0.5 to 0.95).

Experiment 1 served as the baseline, yielding an mAP@50 of 0.649 and an mAP@50–95 of 0.391. The optimal configuration considering both metrics is D=5/N=7. The model using the configuration achieved improvements of 2.5% and 3.9% in mAP@50 and mAP@50–95, respectively. Fig. 4 presents the training convergence curves, with subplots (a) and (b) reporting mAP@50 and mAP@50–95, respectively. Collectively, these results confirmed that the D5N7 configuration achieves the optimal balance of mAP@50 and mAP@50–95. The PR curves for each model experiment are shown in Fig. 5.

We outputed the data from the 17th layer of the entire network, and generated a heatmap as shown in Fig. 6. The heatmap generated by RRT-YOLO showed better recognition and localization of pests than that generated by the original YOLOv12.

3.4 Comparison Results with Other Models

To validate the effectiveness of our proposed model, we conducted a comprehensive performance comparison between

Experiment	Parameters (M)	GFLOPs	D	N	mAP@50	mAP@50–95
1	2.53	6.0	/	/	0.649	0.391
2	2.53	6.0	3	3	0.649	0.403
3	2.53	6.0	5	5	**0.679**	0.417
4	2.53	6.0	7	7	0.654	0.395
5	2.53	6.0	9	9	0.647	0.397
6	2.53	6.0	7	5	0.653	0.408
7	2.53	6.0	5	7	0.674	**0.43**

Table 1 Comparison of experimental results under different D/N parameters on the IP102 dataset.

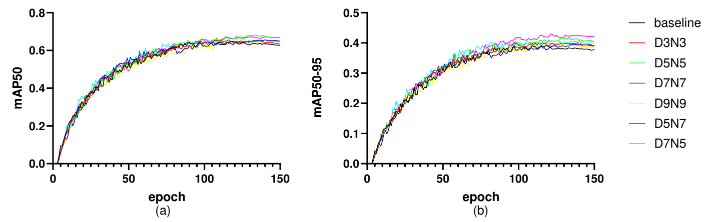

Fig. 4 Comparison of algorithm detection results on the IP102 dataset. (a) Curves of mAP@50; (b) Curves of mAP@50–95.

RRT-YOLO and a range of state-of-the-art object detection models, including two-stage detectors (Faster RCNN), single-stage detectors (SSD, YOLOv8n–YOLOv13n), and Transformer-based detectors (DETR variants), as shown in Table 2. The evaluation metrics included model size (parameters, in M), computational complexity (GFLOPs), and detection accuracy (mAP@50 and mAP@50–95).

Among the compared models, Faster RCNN (resnet50) exhibited high computational overhead (470.7 GFLOPs) and large parameter size (28.6 M), yet achieves only moderate accuracy. SSD (vgg) also suffered from high complexity (64.2 GFLOPs) and large parameters (28.29 M). While lightweight YOLO variants (e.g., YOLOv8n, YOLOv10n–YOLOv13n) reduced model size and computation. Transformer-based DETR, despite larger parameter sizes and higher GFLOPs, failed to outperform YOLO models in accuracy. Notably, RRT-YOLO got the best overall trade-off between efficiency and accuracy. Although YOLOv8n achieved the highest mAP@50, RRT-YOLO achieved the highest mAP@50–95. More importantly, compared to YOLOv8n, RRT-YOLO had 16.2% fewer parameters and 29.4% less computational cost.

3.5 Results on Tomato Leaf Diseases Datasets

The tomato leaf diseases detection dataset is a comprehensive object detection resource publicly hosted on the Roboflow Universe platform [22]. The dataset consists of a total of 737 images, including 645 in the training set, 61 in the validation set, and 31 in the test set. Seven leaf disease categories are included, namely Bacterial Spot, Early Blight, Healthy, Late Blight, Leaf Mold, Target Spot, and Black Spot. Given the relatively small size of this dataset, we conducted training for 200 epochs. As shown in Table 3, a series of experiments were conducted on the tomato

Fig. 5 PR curves of YOLOv12n and RRT-YOLO on the IP102 validation dataset. (a) PR curve of YOLOv12n; (b) PR curve of RRT-YOLO.

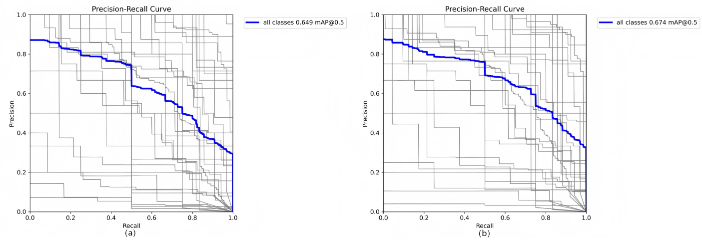

Table 2 Comparison of detection metrics between RRT-YOLO and other advanced models on the IP102 dataset.

Model	Parameters (M)	GFLOPs	mAP@50	mAP@50–95
FasterRCNN(ResNet50)	28.6	470.7	0.638	0.326
SSD (VGG)	28.29	64.2	0.618	0.378
YOLOv8n	3.02	8.2	**0.697**	0.426
YOLOv10n	2.72	8.5	0.665	0.415
YOLOv11n	2.60	6.5	0.685	0.418
YOLOv12n	2.53	**6.0**	0.649	0.391
YOLOv13n	**2.47**	6.4	0.65	0.407
DETR (L)	32.9	108.1	0.547	0.329
DETR (ResNet50)	42.8	130.6	0.617	0.387
RRT-YOLO	2.53	**6.0**	0.674	**0.43**

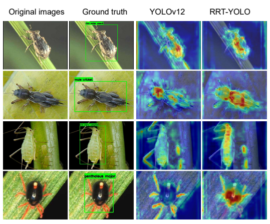

Fig. 6 Comparison of YOLOv12n and RRT-YOLO heatmaps of four insects.

dataset to investigate the impact of D (magnitude of random rotation along the channel dimension) and N (magnitude of random rotation along the sequence dimension) on detection performance, with model parameters and computational complexity held constant across all experiments. The optimal configuration (D=7, N=5) improved mAP@50 by 8.8% and mAP@50–95 by 4.2% compared to baseline.

Fig. 7 presented the training convergence curves of mAP metrics for YOLOv12 (baseline) and four of its variants. Across both evaluation metrics, the D7N5 variant emerged as the optimal model. The PR curves are shown in Fig. 8.

4 Discussion

This study improved the Transformer by introducing rotation in the channel and sequence dimensions, thereby enhancing detection accuracy. On the IP102 dataset, mAP@50 was improved by 2.5% to 67.4%, and mAP@50–95 was improved by 3.9% to 43%.

LP-YOLO [23] introduced lightweight components, namely `LP_Unit` and `LP_DownSample`. C3M-YOLO [24] performed data augmentation on IP102 using Mosaic enhancement and the proposed C3M module flexibly handled

Table 3 Comparison of experimental results under different D/N parameters on the tomato leaf diseases dataset.

Experiment	Parameters (M)	GFLOPs	D	N	mAP@50	mAP@50–95
1	2.52	6.0	/	/	0.597	0.414
2	2.52	6.0	5	5	0.639	0.424
3	2.52	6.0	5	7	0.651	0.432
4	2.52	6.0	7	5	**0.685**	**0.456**
5	2.52	6.0	7	7	0.597	0.395

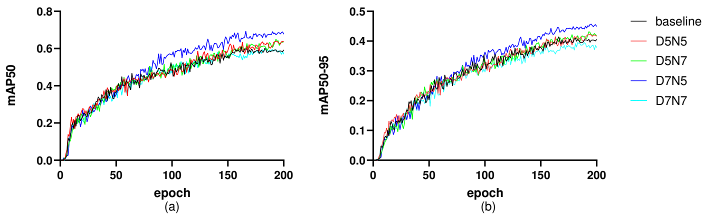

Fig. 7 Comparison of algorithm detection results on the tomato leaf diseases dataset. (a) Curves of mAP@50; (b) Curves of mAP@50–95.

image features while improving inference speed. AEC-YOLOv8n [25] proposed the AFEM-SIE and EMSFEM to effectively handle complex features and high inter-class similarity. The experimental comparisons of the above models were shown in Table 4. Compared to the other three models, RRT-YOLO's drawback is that it does not reduce the model size, but RRT-YOLO achieves better accuracy. RRT-YOLO can be further improved in the future to reduce model size and computational cost.

5 Conclusions

In this study, we introduced an enhancement for Transformer, namely Random Rolling Transformer (RRT). Specifically, we embedded RRT into the Area Attention of YOLOv12n. During training, the query (Q), key (K), and value (V) tensors were randomly rolled along the channel and sequence dimensions before computing attention. This operation increased the diversity of attention patterns and encouraged the network to exploit neighboring channel and spatial information.

To evaluate the effectiveness of this design, we selected 36 pest categories related to rice, wheat, and maize (the main crops in our target deployment area, Henan Province) from the publicly available IP102 dataset. We trained the YOLOv12 baseline and our RRT-YOLO model under the same settings and compared their detection performance. To assess generalization, we further validated the approach on a tomato leaf disease dataset. Experimental results showed that RRT-YOLO achieved consistent improvements while maintaining the same number of parameters and computational complexity: mAP@0.5 and mAP@0.5–0.95 were improved by 2.5% and 3.9% on IP102, and by 8.8% and 4.2% on the tomato leaf disease dataset. Ablation studies indicated that both channel and sequence rolling contributed to the performance gains. In addition, compared with other advanced models such as YOLOv8/9/10 and DETR, RRT-YOLO achieved a better trade-off between accuracy and efficiency.

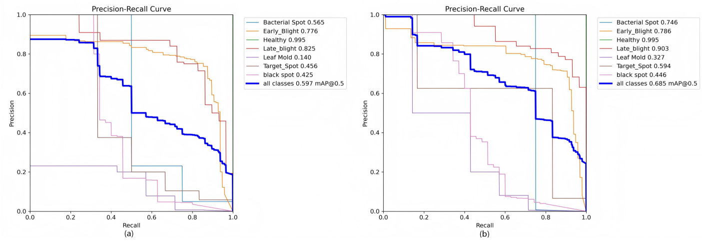

Fig. 8 PR curves of YOLOv12n and RRT-YOLO on the tomato leaf disease validation dataset.

Table 4 Comparison of RRT-YOLO and three other models.

Model	Parameters (M)	GFLOPs	mAP@50	mAP@50–95
LP-YOLO	**1.00**	**3.4**	0.622	0.398
C3M-YOLO	7.12	16.1	0.572	0.349
AEC-YOLOv8n	7.80	/	0.671	**0.431**
RRT-YOLO	2.52	6.0	**0.674**	0.43

The Random Rolling Transformer proposed in this paper is a generic attention perturbation strategy and may benefit other Transformer-based tasks. In future, we will explore its application to other domains.

Declarations

- Funding
  This study was supported by the Key Research and Promotion Special Project of Xuchang City in 2025 (No. 2025090), the 2026 Key Scientific Research Projects of Higher Education Institutions in Henan Province (No. 26B460032), and the Young Key Teacher Training Program of Zhongyuan Institute of Science and Technology (No. ZYKJQNGG2510).

- Declaration of generative AI and AI-assisted technologies in the manuscript preparation process
  During the preparation of this work, the authors used Prism (GPT-5.2) for English spelling, grammar checking, and formatting generation. After using this service, the authors reviewed and edited the content as needed and took full responsibility for the content of the published article.
- Data availability
  The datasets used in this paper are publicly available and can be downloaded from the Internet.
- Conflict of interest
  The authors have no conflict of interest to declare that is relevant to the content of this article.

Contributions

XD involved in conceptualization, investigation, writing-original draft. WH took part in methodology, software, writing-review, editing. YD and AZ involved in data curation, formal analysis, validation. JK took part in experiment implementation, visualization.

References

[1] Chin, P.-W., Ng, K.-W., Palanichamy, N.: Plant disease detection and classification using deep learning methods: A comparison study. Journal of Informatics and Web Engineering **3**(1), 155–168

(2024) https://doi.org/10.33093/jiwe.2024.3.1.10

[2] Wang, Y., Yi, C., Huang, T., Liu, J.: Research on intelligent recognition for plant pests and diseases based on improved yolov8 model. Applied Sciences **14**(12), 5353 (2024) https://doi.org/10.3390/app14125353

[3] Dong, Q., Sun, L., Han, T., Cai, M., Gao, C.: Pestlite: A novel yolo-based deep learning technique for crop pest detection. Agriculture **14**(2), 228 (2024) https://doi.org/10.3390/agriculture14020228

[4] Xiang, Q., Huang, X., Huang, Z., Chen, X., Cheng, J., Tang, X.: Yolo-pest: An insect pest object detection algorithm via cac3 module. Sensors **23**(6), 3221 (2023) https://doi.org/10.3390/s23063221

[5] Liu, H.-I., Tseng, Y.-W., Chang, K.-C., Wang, P.-J., Shuai, H.-H., Cheng, W.-H.: A denoising fpn with transformer r-cnn for tiny object detection. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–15 (2024) https://doi.org/10.1109/TGRS.2024.3396489

[6] Meng, H., Tan, J.: An improved algorithm for small target detection based on yolo. Frontiers in Computing and Intelligent Systems **10**(3), 18–22 (2024) https://doi.org/10.54097/e6sq4903

[7] Kim, G.-E., Son, C.-H.: Roi-aware multiscale cross-attention vision transformer for pest image identification. Computers and Electronics in Agriculture **237**, 110732 (2025) https://doi.org/10.1016/j.compag.2025.110732

[8] Zhang, J., Liu, Z., Yu, K.: MSFNet-CPD: Multi-Scale Cross-Modal Fusion Network for Crop Pest Detection. arXiv preprint. arXiv:2505.02441 (2025). https://doi.org/10.48550/arXiv.2505.02441

[9] Xiong, J., Wu, J., Tang, M., Xiong, P., Huang, Y., Guo, H.: Combining YOLO and background subtraction for small dynamic target detection. The Visual Computer **41**(1), 481–490 (2025) https://doi.org/10.1007/s00371-024-03342-1 . Accessed 2026-04-15

[10] Zheng, Y., Zheng, W., Du, X.: A lightweight rice pest detection algorithm based on improved yolov8. Scientific Reports **14**, 29888 (2024) https://doi.org/10.1038/s41598-024-81587-5

[11] Yuan, W., Lan, L., Xu, J., Sun, T., Wang, X., Wang, Q., Hu, J., Wang, B.: Smart agricultural pest detection using i-yolov10-sc: An improved object detection framework. Agronomy **15**(1), 221 (2025) https://doi.org/10.3390/agronomy15010221

[12] Wang, J., Wang, T., Xu, Q., Gao, L., Gu, G., Jia, L., Yao, C.: Rp-detr: end-to-end rice pests detection using a transformer. Plant Methods **21**, 63 (2025) https://doi.org/10.1186/s13007-025-01381-w

[13] Hao, S., Gao, E., Ji, Z., Ganchev, I.: Bcs_yolo: Research on corn leaf disease and pest detection based on yolov11n. Applied Sciences **15**(15), 8231 (2025) https://doi.org/10.3390/app15158231

[14] Lin, C., Zou, C., Xu, H.: Scnet:

A dual-branch network for strong noisy image denoising based on swin transformer and convnext. Computer Animation and Virtual Worlds **36**(3), 70030 (2025) https://doi.org/10.1002/cav.70030 . Accessed 2026-04-15

[15] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS) (2017)

[16] Ma, P., He, X., Chen, Y., Liu, Y.: ISOD: improved small object detection based on extended scale feature pyramid network. The Visual Computer **41**(1), 465–479 (2025) https://doi.org/10.1007/s00371-024-03341-2 . Accessed 2026-04-15

[17] Tian, Q., Wu, Y., Zhou, Y., Wang, Z., Ye, Q., Huang, G., Li, X.: YOLOv12: Attention-Centric Real-Time Object Detectors. arXiv preprint. arXiv:2502.12524 (2025). https://doi.org/10.48550/arXiv.2502.12524

[18] Ultralytics: YOLOv12 documentation. Online. https://docs.ultralytics.com/models/yolo12/ Accessed 2026-04-15

[19] Dao, T., Fu, D.Y., Ermon, S., Rudra, A., Ré, C.: Flashattention: Fast and memory-efficient exact attention with IO-awareness. In: Advances in Neural Information Processing Systems (NeurIPS) (2022)

[20] Wu, X., Zhan, C., Lai, Y., Cheng, M.-M., Yang, J.: Ip102: A large-scale benchmark dataset for insect pest recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)

[21] Kulyk, L.: IP102 YOLOv5 (Kaggle dataset). Kaggle. https://www.kaggle.com/datasets/leonidkulyk/ip102-yolov5 Accessed 2026-04-15

[22] Sylhet Agricultural University: Tomato Leaf Diseases Detect (Roboflow Universe dataset). Roboflow Universe. https://universe.roboflow.com/sylhet-agricultural-university/tomato-leaf-diseases-detect/dataset/3 Accessed 2026-04-15

[23] Yu, J., Qiu, B., Song, Q., Xu, W., Wang, L., Lian, X., Chen, G.: Lp-yolo: A lightweight model for crop pest detection on ip102 dataset. Agriculture **14**(8), 1420 (2024) https://doi.org/10.3390/agriculture14081420

[24] Xiao, W., Chen, Z., Xu, Y., *et al.*: C3m-yolo: An improved yolov5 model for pest detection on the ip102 dataset. Agronomy **13**(6), 1583 (2023) https://doi.org/10.3390/agronomy13061583

[25] Wei, J., Ni, Q., Chen, Q., Li, C., Jiang, X., Chen, F., Deng, X.: Aec-yolov8n: An efficient model for pest detection based on attention enhancement and cross-scale feature fusion. Agronomy **14**(8), 1640 (2024) https://doi.org/10.3390/agronomy14081640