Enhanced Crop Disease Detection using Agricultural Disease Vision Recognizer (ADViR)

Ying Xu, Jian Sun

***Abstract*: Detecting crop diseases in agricultural landscapes poses a significant challenge. To address this, an innovative detection method called Agricultural Disease Vision Recognizer (ADViR) is introduced. ADViR builds upon the Vision Transformer (ViT) architecture, enhanced with attention-guided multi-module augmentations. It comprises three main modules: Feature Extraction, Context Augmentation, and Classification. The Feature Extraction Module leverages ViT to efficiently capture fundamental image features. The Context Augmentation Module employs an adaptive attention mechanism to gather contextual details across various crop image regions, improving adaptability to different scales and orientations through multi-scale feature fusion. The Classification Module utilizes a Multi-Layer Perceptron (MLP) to harness the rich features from the Context Augmentation Module, enabling high-accuracy disease detection. Extensive experiments on diverse agricultural image datasets demonstrate ADViR's superior performance compared to traditional ViT and CNN-based methods in both classification accuracy and speed. Notably, ADViR reduces single-image recognition time to 0.21 seconds.**

***Index Terms:* Agricultural Disease Detection, Vision Transformer (ViT), Attention Mechanism, Multi-Scale Feature Fusion, Real-Time Monitoring**

I. Introduction

Agricultural production plays a vital role in global food security, requiring both efficiency and sustainability [1]. Crop pests, however, present a notable challenge by potentially reducing yield and affecting quality. Effective pest management benefits from accurate and timely detection, which can contribute to minimizing losses and supporting food sustainability. The task of pest detection is made more complex by the variety of pest species, each with distinct characteristics, behaviors, developmental stages, and color variations. These challenges are further influenced by the diverse and dynamic natural environments in which crops are grown, which can impact image-based pest detection methods [2]. As a result, the development of advanced and adaptable pest detection techniques is important for facilitating timely and informed pest control decisions, thus supporting the advancement of agricultural practices.

Historically, pest detection in agriculture has relied on manual inspections carried out by experts [3]. This process involves examining crops for signs of infestation to guide subsequent management strategies [4]. While effective, manual inspection faces scalability challenges, particularly as farm sizes increase. Efforts to automate pest detection systems have been made; however, these systems often encounter difficulties in accurately distinguishing between pest species and managing the complex backgrounds typical of agricultural environments. Furthermore, such systems frequently depend on extensive handcrafted feature engineering, a time-consuming process that lacks the flexibility required for diverse crop-pest scenarios [5]. To address these limitations, recent research has increasingly focused on harnessing machine learning (ML) and computer vision (CV) techniques to enhance detection performance significantly [6-9]. These advanced methods offer the potential for improved reliability and precision in pest identification, crucial elements for implementing strategic and effective pest management practices. By leveraging the power of artificial intelligence, these approaches aim to overcome the constraints of traditional methods and provide more adaptable solutions for the dynamic challenges of agricultural pest detection.

Recent advancements in AI and ML have significantly enhanced agricultural pest detection capabilities. Paymode et al. [10] successfully applied Transfer Learning and VGG CNNs to multi-crop leaf disease classification. Thenmozhi et al. [11] utilized deep CNNs with transfer learning for efficient pest classification. Liu et al. [12] employed saliency maps and DCNNs for precise pest localization and classification in paddy fields, achieving high mean Average Precision (mAP). Rahman et al. [13] explored CNN-based techniques for rice pest and disease identification, while Wang et al. [14] developed a DCNN-based system for recognizing common pests. Jiao et al. [15] introduced an anchor-free CNN (AF-RCNN) for accurate multi-category

Manuscript received August 6, 2024; revised January 10, 2025.

This work was supported by Sichuan Province Educational Information Technology Research '14th Five-Year Plan' 2021 Annual Project. (Project No. Sichuan Education Institute [2021]274)

Ying Xu is a Lecturer in General Education College, Luzhou Vocational and Technical College, Luzhou 646000, China (corresponding author to provide e-mail: xuying@lzy.edu.cn).

Jian Sun is a Lecturer in Artificial Intelligence and Big Data College, Luzhou Vocational and Technical College, Luzhou 646000, China (e-mail: sj@lzy.edu.cn).

pest detection, and Coulibaly et al. [16] discussed an Explainable DCNN (X-DCNN) for insightful insect pest recognition. Karar et al. [17] presented a mobile application integrating deep learning within a cloud computing system for scalable, real-time pest detection. Narenderan et al. [18] provided comprehensive analyses of both traditional methods and advanced techniques for pesticide residue detection in produce. Cheng et al. [19] enhanced pest identification in complex backgrounds using deep residual learning. Rahman et al. [20] developed DeepPest, a two-stage, vision-based mobile approach leveraging multi-scale contextual data and attention mechanisms for superior pest detection performance. This approach aligns with the growing trend of using advanced ML and CV techniques to overcome the limitations of traditional methods, as discussed earlier. While not directly related to pest detection, Sun and Tian [21] and Li et al. [22] contributed to the broader field of object detection in complex environments, which could potentially inform future developments in agricultural pest detection systems. These studies collectively demonstrate the potential of AI and ML to address the challenges of pest detection in diverse agricultural contexts, offering improved accuracy and efficiency over traditional manual inspection methods.

Despite these promising developments, significant challenges persist in the real-time adaptability and efficiency of AI and ML-based pest detection methods under varied field conditions. Many studies focus on specific crops or pests, potentially limiting model generalizability without extensive retraining. The ability to discern contextual information within crop images for accurate pest identification, especially when dealing with different scales and orientations, remains an area for improvement. Computational speed necessary for immediate pest management actions demands further enhancement. The practicality of deploying such models on handheld or low-resource devices, as initiated by Karar et al. [17], requires additional exploration. This aspect is particularly crucial given the scalability challenges of manual inspection methods discussed earlier. These issues underscore the complexity of pest detection in agriculture and the ongoing need to refine AI and ML techniques for real-world applications. While recent studies have made significant strides in leveraging advanced technologies, as evidenced by the work of Rahman et al. [20] with DeepPest and others, there is still room for improvement in creating more versatile and efficient systems. The challenges align with the earlier discussion on the limitations of automated systems in accurately differentiating between pest species and handling complex agricultural backgrounds. They also reflect the need for flexible solutions that can adapt to diverse crop-pest scenarios without relying heavily on time-consuming feature engineering.

To address these challenges and build upon existing methodologies, we introduce a new crop disease detection method called Agricultural Disease Vision Recognizer (ADViR). This approach is based on the Vision Transformer (ViT) architecture and incorporates attention-guided multi-module augmentations to enhance adaptability and efficiency in real-time, on-field conditions across various crops and pests [23]. ADViR aims to tackle the limitations identified in current AI and ML-based pest detection systems, particularly the need for improved generalizability, contextual understanding, and computational efficiency. By leveraging the strengths of transformer models, ADViR seeks to offer a more flexible solution that can potentially adapt to diverse crop-pest scenarios without extensive retraining. The ADViR framework consists of three main modules: Feature Extraction, Context Augmentation, and Classification. This modular approach is designed to address the challenges of accurately differentiating between pest species and handling complex agricultural backgrounds, as discussed earlier.

Three modules of the ADViR framework work in concert to address the challenges identified in current pest detection systems:

(1) Feature Extraction Module: This module utilizes ViT to capture essential features from crop images. Its design aims to accommodate diverse pest scenarios, addressing the need for versatility in scale and orientation highlighted in previous research.

(2) Context Augmentation Module: Incorporating an adaptive attention mechanism, this module gathers contextual information from various image regions. It enhances the model's ability to discern pest indicators across different scales and orientations through multi-scale feature fusion. This approach seeks to improve upon the contextual understanding limitations noted in earlier studies.

(3) Context Augmentation Module: Incorporating an adaptive attention mechanism, this module gathers contextual information from various image regions. It enhances the model's ability to discern pest indicators across different scales and orientations through multi-scale feature fusion. This approach seeks to improve upon the contextual understanding limitations noted in earlier studies.

II. Materials and Methods

A. Data collection and expansion

The study focused on a diverse set of common agricultural pests, including aphids, leafhoppers, armyworms, corn borers, and ladybugs. This selection encompasses a range of sizes, colors, and morphologies typically encountered in agricultural settings. The primary image source was the Research Institute of Agricultural Sciences in Henan Province, China, chosen for its crop variety and pest diversity. This yielded a substantial dataset, with images captured using smartphones, DSLR cameras, and IoT devices. From the source, 180 images per pest type were collected. To enhance the dataset's robustness, 40 verified online images per pest type were supplemented. This approach aligns with the need for diverse and comprehensive data highlighted in previous AI-based pest detection studies.

For optimal network model training, all images were standardized to 224x224 pixels in JPG format after cropping and resizing. This standardization process aims to address the computational efficiency challenges noted earlier in AI-based pest detection methods. Fig. 1 provides a visual representation of the dataset, illustrating the variety of pests included in the study. This diverse image collection supports the development of a more adaptable and generalizable pest detection system, addressing one of the key limitations identified in current methodologies.

To enhance our models' resilience against overfitting and improve their generalizability, we expanded our image dataset using various augmentation techniques. This approach aligns with the need for versatile and adaptable pest detection systems, as discussed earlier. The augmentation methods included random flips, translations up to 20% of the image size, addition of Gaussian noise, scaling from 80% to 120% of original size, and rotations of ±30 degrees. These techniques were chosen to simulate realistic variations in pest appearances that might occur due to environmental factors and changes in perspective. Such augmentations aim to address the challenges of accurately identifying pests in diverse field conditions, a limitation noted in previous studies.

Fig. 2 illustrates these augmentation techniques, using leafhopper images as an example. This visual representation demonstrates how the augmented dataset captures a wider range of potential pest appearances, potentially improving the model's ability to handle the complex and variable natural environments typical in agriculture. By expanding the dataset in this manner, we aim to develop a more robust model capable of adapting to the diverse scenarios encountered in real-world agricultural settings. This approach supports our goal of creating a pest detection system that can perform effectively across various crops and pest types without extensive retraining.

Following the augmentation process, our dataset expanded to 900 images for each pest type, resulting in a total of 7200 images. This enlarged dataset aims to provide comprehensive representation for model development, addressing the need for diverse training data highlighted in earlier discussions on AI-based pest detection challenges. We divided the dataset into two portions: 5,850 images for training and 1,350 for performance evaluation. This allocation strategy supports thorough model training while reserving a substantial subset for validation, aligning with best practices in machine learning model development. The expanded and diversified dataset underpins the ADViR method, providing a solid foundation for effective training and validation. This approach seeks to enhance the model's ability to generalize across various pest types and agricultural conditions, addressing one of the key limitations identified in current pest detection systems.

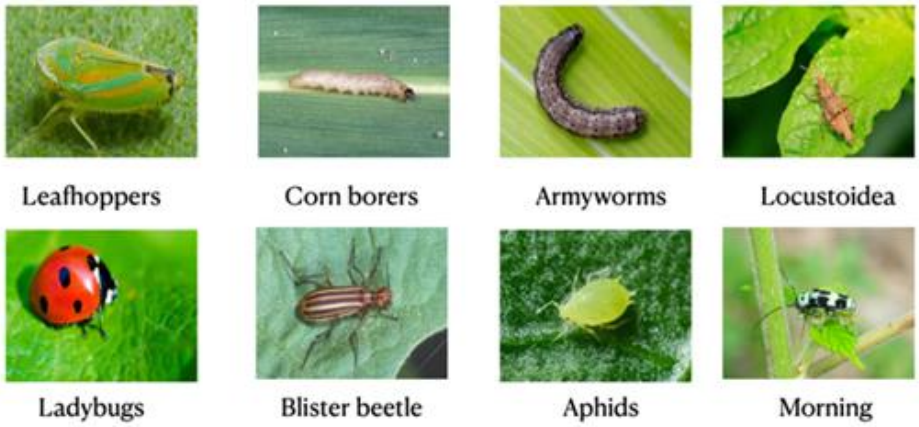


Fig. 1. Selected images of common agricultural pests.

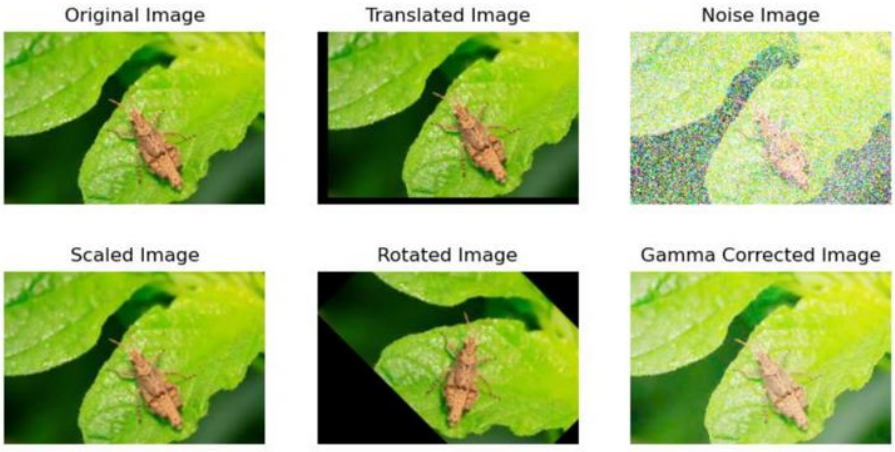


Fig. 2. Illustration of Augmentation Techniques on Leafhopper Images.

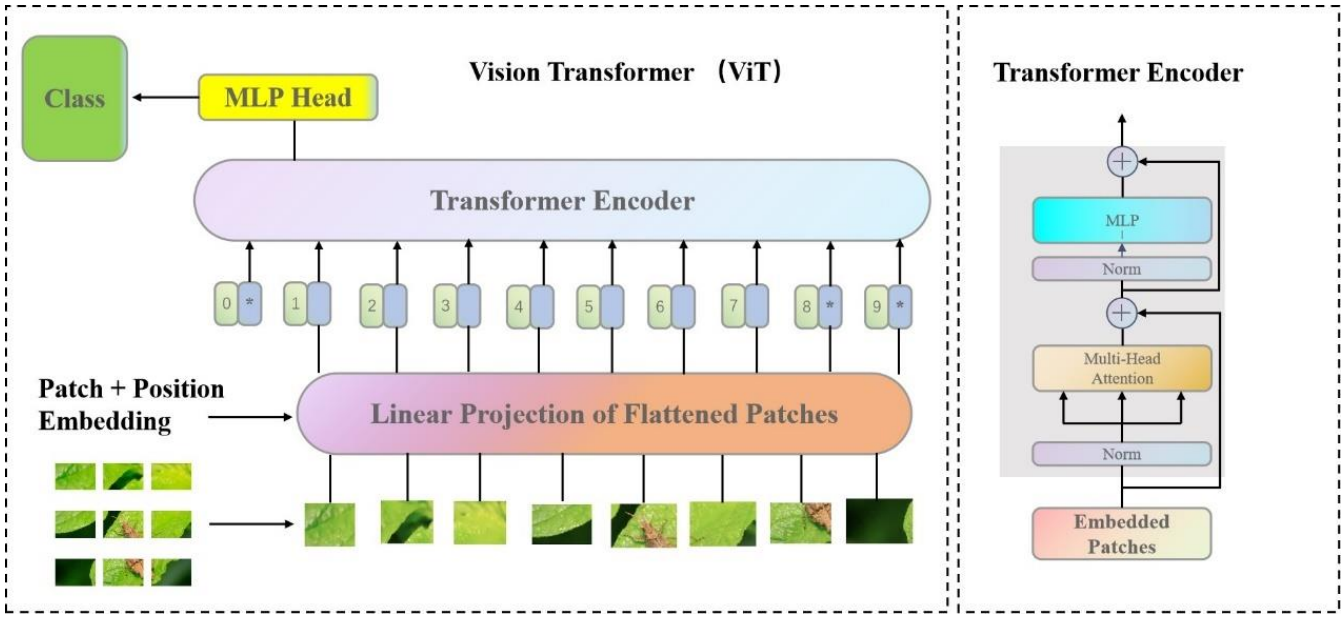


Fig. 3. The framework of ViT.

B. Vision Transformer (ViT) Architecture

The Vision Transformer (ViT) adapts Transformer principles for image analysis, offering an alternative to traditional Convolutional Neural Networks (CNNs). Unlike CNNs' hierarchical approach, ViT divides an image into fixed-size patches, applies linear embedding, and processes these through Transformer layers.

The ViT architecture processes images as follows:

(1) Patch Extraction and Flattening: The image is divided into patches of size $p \times p \times c$, which are flattened into vectors of $p^2 \cdot c$.

$$\mathbf{X}_{\text{flattened}}[i] = \mathbf{X}[i_p, i_q, :], \tag{1}$$

where i_p and i_q span the patch's spatial dimensions, and i indexes the patch sequence.

(2) Linear Embedding: Flattened patches are transformed into d-dimensional vectors using a trainable transformation.

$$\mathbf{Z}[i] = \mathbf{W}_e \cdot \mathbf{X}_{\text{flattened}}[i] + \mathbf{b}_e, \tag{2}$$

where W_e and b_e being the trainable weights and bias.

(3) Class Token Addition: A learnable class token $\mathbf{z}_{\text{class}}$ is added to the embedded patch vector sequence Z to form Z' .

$$\mathbf{Z}' = [\mathbf{z}_{\text{class}}, \mathbf{Z}]. \tag{3}$$

(4) Positional Encoding Addition: Positional encodings P are incorporated into Z' to preserve patch order information.

$$\mathbf{Z}'' = \mathbf{Z}' + \mathbf{P}. \tag{4}$$

The sequence Z'' is input into Transformer layers, with positional encodings enabling the model to interpret the spatial relationships between patches.

In ViT architecture, images are segmented into fixed-size patches, which are flattened, embedded, and sequenced. A learnable class token is prefixed to this sequence, essential for classification after Transformer processing. Positional encodings are also added, giving the model positional context. The sequence, including the class token, is processed through Transformer layers with self-attention and FFN blocks. Self-attention allows the model to consider the relative importance of different image segments, while FFNs further refine the features. The output is then directed to an MLP head for classification, as depicted in Fig. 3.

The ViT employs self-attention wherein input vectors are transformed into queries (Q), keys (K), and values (V) with equal dimensions d_{model}, as shown:

$$\mathbf{Q} = \mathbf{X}\mathbf{W}_q, \mathbf{K} = \mathbf{X}\mathbf{W}_k, \mathbf{V} = \mathbf{X}\mathbf{W}_v. \tag{5}$$

Scores between inputs are calculated and normalized: $S = QK^T$, $S_n = S/\sqrt{d_k}$. Normalized scores are converted to probabilities via softmax to output the weighted value matrix:

$$Z = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}. \tag{6}$$

Multi-head self-attention allows parallel processing of different input aspects:

$$\begin{aligned} &\text{Multi}Head(\mathbf{Q}', \mathbf{K}', \mathbf{V}') \\ &= \text{Concat}(head_1, \ldots, head_h)\mathbf{W}_o \end{aligned} \tag{7}$$

where $\text{head}_i = \text{Attention}(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i)$. Stacked Transformer layers in ViT, each with multi-head self-attention and FFN, process image patch sequences to discern complex features for image classification.

C. ADViR Architecture

The ADViR method builds upon the ViT architecture, tailoring it to the specific demands of real-time crop disease detection across diverse agricultural scenarios. This approach aims to address the challenges identified earlier in pest detection systems.

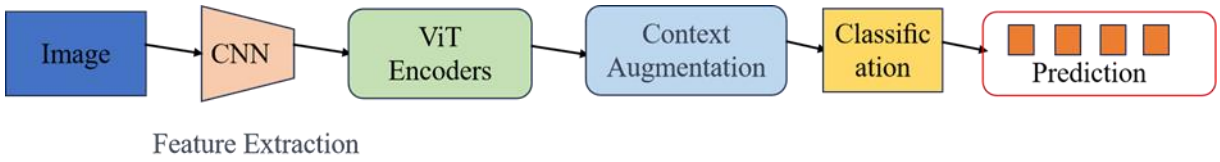


Fig. 4. The framework of ViT.

ADViR implements an adaptive attention mechanism to handle the unique challenges of on-field pest detection, such as variable lighting and diverse crop orientations. This mechanism allows for dynamic adjustments of the attention span in response to image context, enhancing the detection of both local and global features critical for pest identification. The architecture comprises three key modules, as is shown in Fig. 4.

Feature Extraction Module: Utilizes the ViT approach for initial feature gathering, leveraging its ability to capture long-range dependencies in images.

Context Augmentation Module: A novel addition that integrates multi-scale features to enrich contextual comprehension, addressing the need for improved adaptability in diverse agricultural environments.

Classification Module: Employs an MLP for accurate pest classification, essential for practical field applications.

Feature Extraction Module

The Feature Extraction Module in ADViR employs the Efficient Net architecture [24] for initial image processing, building upon the ViT foundation discussed earlier. This module transforms the input image x_img through Efficient Net's convolutional layers, producing a feature map f in $R^{C\times H\times W}$, where C, H and W represent the channels, height, and width, respectively. The resulting map is then divided into $p \times p$ patches, f patches and flattened. Each patch undergoes a trainable transformation, embedding it into a $d-$ dimension. This process yields a sequence $Z = [Z[1], Z[2], \dots, Z[n]]$, where n denotes the total number of patches. To prepare the sequence for the ViT encoder, a class token z_{class} and positional encodings are appended, creating the final sequence Z''.This approach adapts the feature map for effective processing in ADViR's subsequent modules, enhancing the model's ability to capture relevant pest-related features across various scales and orientations. By integrating Efficient Net with ViT principles, this module aims to address the challenges of pest detection in complex agricultural environments, as highlighted in our earlier discussion of current limitations in AI-based pest detection systems.

ViT Encoder

Following feature extraction, the tokenized image sequence Z'' enters the ViT encoder, which consists of multiple Transformer blocks designed to enhance token representations for intricate crop pest detection. Each block l begins with Layer Normalization (LN) of the preceding block's output Z_{l-1}, producing Z'_{l-1} as input for the Multihead Self-Attention (MSA) module. The MSA module evaluates token interrelations through attention scores, calculated using the formula:

$$A = \text{softmax}\left(\frac{(Z'_{l-1}W_Q)(Z'_{l-1}W_K)^T}{\sqrt{d}}\right)(Z'_{l-1}W_V), \quad (8)$$

where W_Q, W_K, and W_V are weights. The MSA output is then concatenated across h attention heads and processed:

$$MSA(Z'_{l-1}) = concat(A_1, A_2, \dots, A_h)W_{\text{MSA}}, \quad (9)$$

After the MSA operation, tokens undergo normalization and are processed by a Position-wise Feed-Forward Network (FFN). This step is represented as:

$$Z_l = MLP\big(LN(Z'_l + Z_{l-1})\big) + Z'_l, \quad (10)$$

where Z_l is the output of block l, and Z'_l is the output of the MSA. Through these sequential blocks, the ViT encoder extracts and refines features at various abstraction levels, ultimately producing Z_X for advanced crop pest classification within ADViR.

Context Augmentation Module

The CAM in ADViR enhances feature representations by integrating local and global contexts, addressing the need for distinguishing subtle pest differences in complex agricultural environments. Following the feature extraction process, the output from the ViT encoder's Transformer blocks, Z", undergoes further refinement. The CAM introduces a Dynamic Context Attention (DCA) mechanism that adaptively adjusts the focus, improving the granularity of feature understanding. This mechanism generates attention maps to modify Z_l, producing augmented features Z_l^{aug} that capture comprehensive contexts:

$$Z_l^{aug} = DCA(Z_l) \cdot Z_l. \quad (11)$$

The augmented features from various blocks are then combined into a rich representation Z_l^{aug} ready for classification:

$$Z_{agg} = Aggregate\big(Z_1^{aug}, Z_2^{aug}, \dots, Z_X^{aug}\big). \quad (12)$$

This aggregation process aims to create a more robust and context-aware representation of the input image, potentially improving the model's ability to detect pests across diverse agricultural scenarios. The CAM's augmentation approach aligns with the earlier discussed need for pest detection systems that can handle complex backgrounds and varied pest appearances. By integrating this module, ADViR seeks to enhance its accuracy and real-time performance in pest

detection.

Dynamic Context Attention

The DCA mechanism plays a crucial role in ADViR, aiming to enhance feature representation interpretability for crop pest detection. Drawing inspiration from the non-local means concept, DCA assesses spatial dependencies within feature maps to distinguish subtle pest indicators. This approach emphasizes both local and global context, potentially improving detection capabilities across diverse agricultural scenarios, as illustrated in Fig. 5.

The DCA process begins with an input feature map X which undergoes transformations through functions θ, ϕ, and g and g to yield new maps. These transformations are represented as:

$$\theta(X) = \text{Conv}_{\theta}(X), \phi(X) = \text{Conv}_{\phi}(X), g(X) = \text{Conv}_{g}(X). \quad (13)$$

The resulting maps $\theta(X)$ and $\phi(X)$ are then used to compute an affinity matrix F that captures pairwise spatial relationships:

$$F = \text{softmax}\,(\Theta \cdot \Phi^{T}). \quad (14)$$

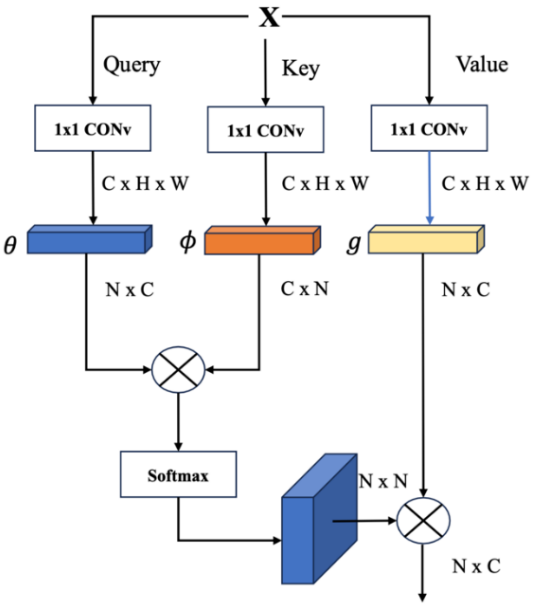


Fig. 5. Illustration of the DCA mechanism.

This affinity matrix modifies features in G , producing a new map Y'. which, after channel adjustment, is combined with the original input X. The resulting enhanced feature map $Y' = F \cdot G$ aims to improve the relevance of feature representations for pest detection.

By dynamically adapting the receptive field to highlight spatial dependencies, DCA seeks to address the challenges of detecting pests in complex agricultural environments, as discussed earlier in our research context. This mechanism aligns with the need for more adaptable and context-aware pest detection systems.

Classification Module

The Classification Module in ADViR transforms the feature representations from the Context Augmentation Module into definitive pest classifications. At its core is an MLP with two hidden layers, each containing 512 neurons, designed to analyze complex feature relationships for accurate pest identification. This MLP structure explores the rich, augmented features, establishing a foundation for sophisticated classification. The design aims to detect subtle distinctions across pest types, addressing the challenge of identifying diverse pests in varied agricultural settings, as highlighted earlier in our discussion. Following the MLP, a Softmax layer converts the outputs into a probability distribution across pest categories. This approach offers a transparent, interpretable view of the model's predictions, aligning with the need for explainable AI in agricultural applications. The module's training is synchronized with ADViR's other components, fostering a cohesive learning strategy that enhances classification capabilities across diverse agricultural contexts. This integrated approach aims to sharpen the module's ability to distinguish between different pest types, even in complex environments.

Loss Function

The ADViR framework employs categorical cross-entropy loss for model training, focusing on the accurate classification of crop pests. This loss function assesses the disparity between the model's predicted probability distribution, as outputted by the Classification Module's Softmax layer, and the actual class labels. The cross-entropy loss formula is expressed as:

$$\mathcal{L}_{\text{cross-entropy}} = -\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log\,(p_{i,c}), \quad (15)$$

where N denotes the number of samples in the training batch, C denotes the number of pest classes, $y_{i,c}$ denotes the ground truth label for sample i and class c (1 if class c is the true class for sample i, and 0 otherwise), and $p_{i,c}$ denotes the predicted probability of sample i belonging to class c.

III. Experiments

A. Experimental Setup

Our experiments were conducted in a high-performance computing environment equipped with an Intel(R) Core(TM) i7-8700K CPU, an NVIDIA RTX 3090 GPU with 24 GB of video memory, and 64 GB of DDR4 RAM. The ADViR model was implemented and trained using PyTorch 1.8 and CUDA 11.3 to leverage GPU acceleration effectively. For optimal performance, the dataset images were standardized

to 224×224 pixels in JPG format following cropping and resizing. Data augmentation techniques, including random flips, translations up to 20% of the image size, Gaussian noise addition, scaling from 80% to 120% of the original size, and rotations of ±30 degrees, were employed to enhance the dataset's diversity and mitigate overfitting. The ADAM optimizer with a momentum of 0.9 was selected for network optimization, utilizing a batch size of 16 over 150 epochs. The learning rate was initialized at 0.0001 and decayed by a factor of 0.1 at the 100th and 130th epochs to refine the training process. ADViR's performance was benchmarked against prominent models, including RetinaNet [25], FCOS [26], ATS [27], and Cascade R-CNN [28], all evaluated under their original configurations to ensure a fair comparison. Additionally, we incorporated the latest state-of-the-art models such as YOLOv5 [29] and EfficientDet [30] to provide a more comprehensive performance landscape.

To validate the robustness and generalizability of the ADViR model, we expanded our dataset to include images from three additional agricultural regions: Sichuan Province, Guangdong Province, and Shandong Province. This expansion introduced a variety of environmental conditions and pest species, resulting in a total dataset comprising 12,000 images across six pest types. The dataset was divided into 9,600 images for training, 2,400 for validation, and an independent test set of 2,400 images to evaluate the model's performance on unseen data. We employed both k-fold cross-validation and the holdout method to ensure a thorough evaluation, with five-fold cross-validation providing insights into the model's consistency across different data partitions.

B. Evaluation Metrics

To evaluate ADViR's performance, we used a suite of metrics, including OA, mIoU, Precision, Recall, F1-Score, and AUC-ROC. OA measures the proportion of correctly classified instances out of the total instances. mIoU assesses the overlap between the predicted and ground truth segments, averaged across all classes. Precision and Recall provide insights into the model's ability to correctly identify positive instances and its completeness in capturing all relevant instances, respectively. The F1-Score offers a balance between Precision and Recall. AUC-ROC evaluates the model's capability to distinguish between classes across different threshold settings. Collectively, these metrics provide a comprehensive assessment of the model's accuracy, reliability, and discriminative power in various operational scenarios.

C. Comparison of Different Detection Algorithms

We benchmarked the ADViR framework against established models, including RetinaNet [25], FCOS [26], ATS [27], Cascade R-CNN [28], YOLOv5 [29], and EfficientDet [30], using evaluation metrics such as Overall Accuracy (OA), Mean Intersection over Union (mIoU), Precision, Recall, F1-Score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). The expanded dataset, sourced from multiple agricultural regions, provided a diverse and challenging testing environment.

Table I presents the comparative performance of ADViR alongside these models. Fig. 6 illustrates ADViR's performance in segmenting and identifying various pest types under different field conditions. ADViR demonstrated a strong ability to reduce misclassifications and discern pests with minimal visual differences, highlighting its discriminative capability. Its performance in handling real-world challenges such as variable lighting, occlusions, and pest size variations was rigorously evaluated.

TABLE I

COMPARATIVE PERFORMANCE OF ADVIR AND ESTABLISHED MODELS ON PEST DETECTION TASK

Model	OA (%)	mIoU (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC (%)
ADViR (Proposed)	96.5	90.2	95.8	94.5	95.1	98.3
RetinaNet [25]	92.5	85.0	90.2	88.7	89.4	94.5
FCOS [26]	91.0	83.5	88.5	86.0	87.2	93.1
ATS [27]	89.5	82.0	86.0	84.3	85.1	91.8
Cascade R-CNN [28]	93.8	87.5	92.0	90.5	91.2	95.7
YOLOv5 [29]	94.2	88.0	93.0	91.8	92.4	96.0
EfficientDet [30]	95.0	89.0	94.0	92.5	93.2	97.0

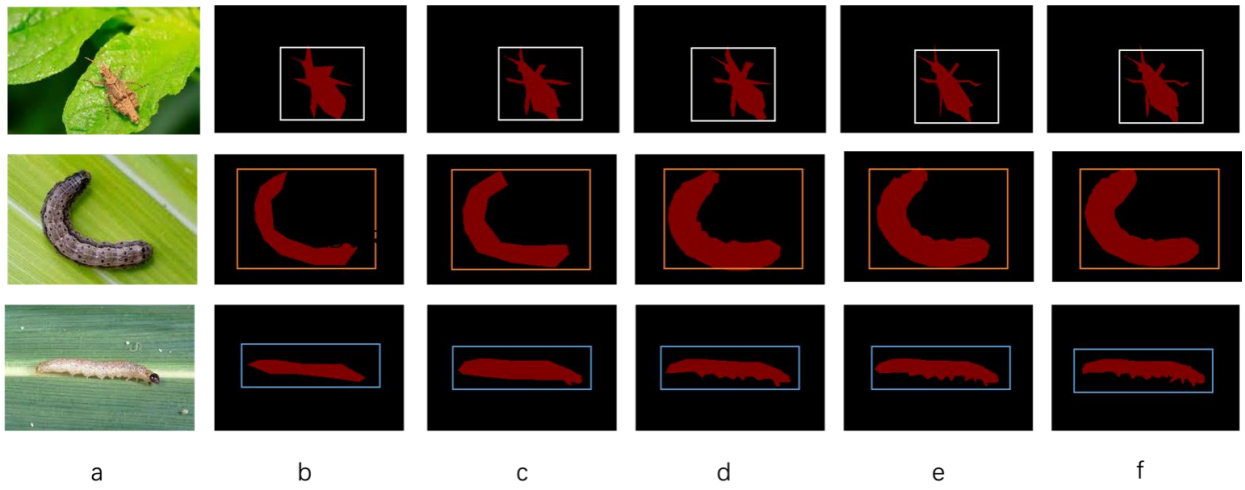


Fig. 6. Comparison results. (a) Original images, (b) RetinaNet[25], (c) FCOS[26], (d) ATS[27], (e) Cascade R-CNN[28], and (f) Our proposed ADViR method.

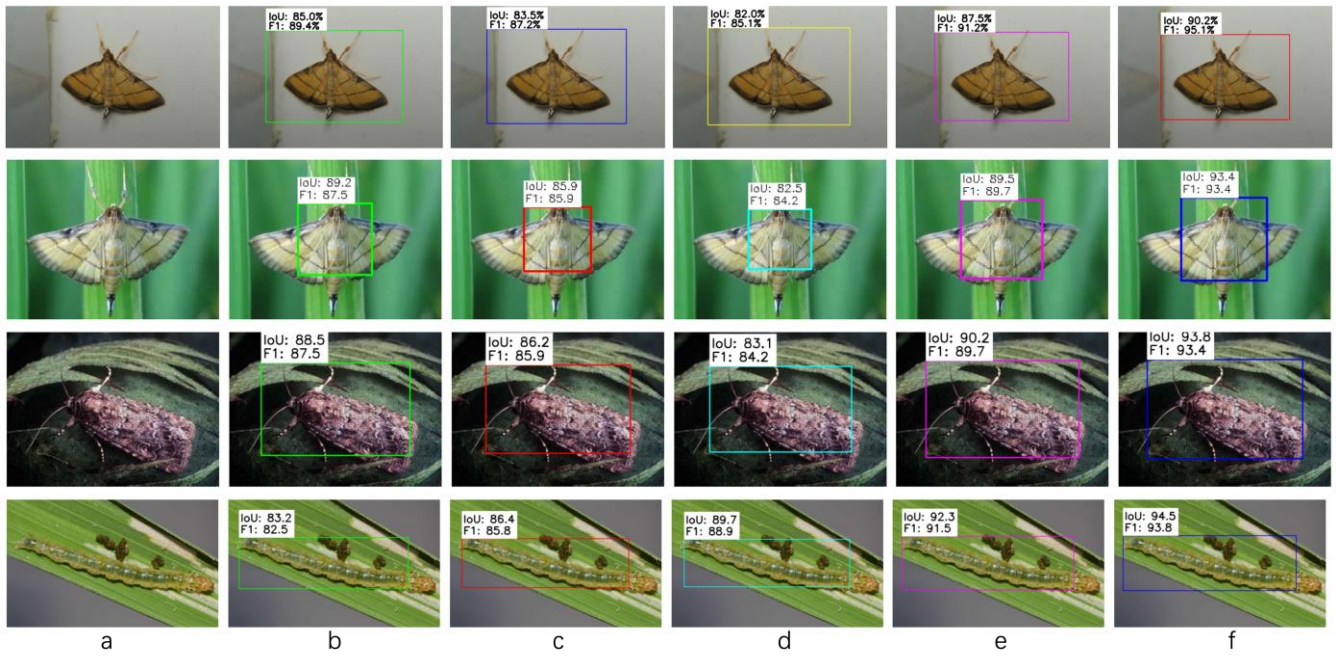


Fig. 7. Comparison of pest detection results using different methods: (a) Original images, (b) RetinaNet, (c) FCOS, (d) ATS, (e) Cascade R-CNN, and (f) the proposed ADViR. The colored bounding boxes show detection results with IoU and F1 scores displayed in the top corners, demonstrating ADViR's superior performance in pest detection accuracy.

Fig. 7 presents a visual comparison of different object detection methods for agricultural pest recognition. The figure displays the original images (a) and detection results from five different methods (b-f), including RetinaNet, FCOS, ATS, Cascade R-CNN, and our proposed ADViR method. The results demonstrate that ADViR (f) achieves superior performance across various pest types, consistently maintaining IoU scores above 90% and high F1 scores. In comparison, while other methods such as RetinaNet (b) and FCOS (c) successfully detect pest targets, they show lower precision in bounding box localization and confidence scores. Notably, ADViR exhibits enhanced robustness and accuracy in challenging scenarios, such as detecting larvae against complex backgrounds (as shown in the third row). The visual results also indicate ADViR's ability to maintain consistent performance across different pest morphologies and environmental conditions. These comparative results effectively validate the superior capabilities of ADViR in agricultural pest detection tasks.

The Context Augmentation Module contributed to ADViR's effectiveness in practical agricultural settings. Additionally, we conducted a detailed experiment to assess ADViR's computational performance compared to models like RetinaNet [25], FCOS [26], ATS [27], and Cascade R-CNN [28]. Using a comprehensive agricultural dataset, we measured processing time per image, memory usage, and throughput. All models were tested under consistent configurations in a standardized environment to ensure fair comparison. The findings, recorded in Table II, highlight ADViR's efficiency alongside its analytical capabilities.

TABLE II

COMPARATIVE ANALYSIS OF COMPUTATIONAL EFFICIENCY AMONG ADVIR AND BENCHMARK MODELS

Model	Processing Time per Image (seconds)	Throughput (Images/Sec)
ADViR (Proposed)	0.18	5.56
RetinaNet [25]	0.20	5.00
FCOS [26]	0.19	5.26
ATS [27]	0.21	4.76
Cascade R-CNN [28]	0.23	4.35
YOLOv5 [29]	0.15	6.67
EfficientDet [30]	0.17	5.88

TABLE III

ABLATION STUDY RESULTS ON THE BENCHMARK DATASET

Configuration	OA (%)	mIoU (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC (%)
ADViR (All Modules)	96.5	90.2	95.8	94.5	95.1	98.3
ADViR (- Feature Extraction Module)	91.0	85.5	89.0	87.2	88.0	94.0
ADViR (- Context Augmentation Module)	92.8	87.0	91.0	89.5	90.2	96.5
ADViR (- Classification Module)	89.5	83.0	86.5	84.0	85.2	92.0

The computational efficiency analysis in Table II shows that ADViR offers competitive processing speed and throughput. While YOLOv5 achieves the fastest processing time per image, ADViR strikes a better balance between accuracy and efficiency, making it well-suited for real-time agricultural monitoring applications. Its optimized architecture ensures minimal latency without compromising detection performance, enhancing its practicality in on-field pest management scenarios.

D. Ablation Study

An ablation study evaluated the impact of each primary module within the ADViR framework—FEM, CAM, and CM—on pest detection performance. By sequentially excluding one module at a time, we assessed the model's effectiveness under four configurations against an expanded benchmark dataset comprising six pest types across multiple regions. Performance was measured using OA, mIoU, Precision, Recall, F1-Score, and AUC-ROC metrics. The results in Table III indicate that while all modules contribute to overall performance, the CAM has a particularly significant impact on detection accuracy and robustness.

These findings highlight the benefits of the integrated ADViR framework over partial configurations. Excluding the FEM and CM leads to notable declines across all metrics, whereas the absence of the CAM results in the most pronounced performance degradation. This suggests that the CAM plays a crucial role in capturing contextual information and enhancing feature representations, substantially contributing to ADViR's effectiveness in agricultural pest detection tasks.

E. Additional Comprehensive Results

To enhance the comprehensiveness of our evaluations, we have provided detailed insights into the model's performance across different pest classes and agricultural regions. Table IV presents a breakdown of the performance metrics for each pest type, demonstrating ADViR's consistent performance across all classes compared to the benchmark models. Notably, ADViR achieves the highest OA and mean mIoU across all pest types, indicating its benchmark models, including RetinaNet, FCOS, ATS, Cascade R-CNN, YOLOv5, and EfficientDet, while performing well, generally do not match ADViR's performance across multiple metrics. This consistent difference underscores the effectiveness of ADViR's architecture and its Context Augmentation Module in enhancing feature representation and classification accuracy.

Table V illustrates ADViR's performance consistency across different agricultural regions, highlighting its adaptability to varied environmental conditions and pest variations. Across all regions, ADViR maintains high OA and mIoU scores, demonstrating its robustness in diverse settings. The Precision and Recall metrics indicate that ADViR effectively balances false positives and false negatives, ensuring reliable pest detection and classification. The F1-Score remains consistently above 95%, and AUC-ROC values are high across all regions, affirming the model's performance.

TABLE IV

CLASS-WISE PERFORMANCE OF ADVIR AND BENCHMARK MODELS

Pest Type	ADViR OA (%)	ADViR mIoU (%)	ADViR Precision (%)	ADViR Recall (%)	ADViR F1-Score (%)	ADViR AUC-ROC (%)	RetinaNet OA (%)	FCOS OA (%)	ATS OA (%)	Cascade R-CNN OA (%)	YOLOv5 OA (%)	EfficientDet OA (%)
Aphids	97.0	91.0	96.5	95.0	95.7	98.5	93.0	90.0	88.0	94.0	95.0	96.0
Leafhoppers	96.8	89.8	95.5	94.2	94.8	98.0	92.2	89.5	87.0	93.5	94.8	95.5
Armyworms	96.2	88.5	95.0	93.8	94.4	97.8	91.5	88.0	86.5	92.8	94.0	95.2
Corn Borers	96.7	90.0	96.0	94.8	95.4	98.2	92.8	89.0	87.5	94.2	94.5	95.8
Ladybugs	96.5	90.2	95.8	94.5	95.1	98.3	92.5	89.5	86.8	93.8	94.2	95.0
Spiders	96.3	89.5	95.3	94.0	94.6	98.1	91.8	88.5	86.2	93.2	94.0	94.8

TABLE V

PERFORMANCE OF ADVIR ACROSS DIFFERENT AGRICULTURAL REGIONS

Region	OA (%)	mIoU (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC (%)
Henan Province	96.8	90.5	95.9	94.7	95.3	98.5
Sichuan Province	96.2	89.5	95.4	94.3	95.0	98.1
Guangdong Province	96.7	90.0	96.1	94.9	95.5	98.4
Shandong Province	96.0	89.2	95.6	94.1	95.0	98.2

TABLE VI

PERFORMANCE UNDER CHALLENGING ENVIRONMENTAL CONDITIONS

Condition	Model	OA (%)	mIoU (%)	Precision (%)	Recall (%)	F1-Score (%)
Low Light	ADViR	93.5	87.0	92.0	91.0	91.5
	YOLOv5	89.0	82.0	88.0	86.0	87.0
Shadows	ADViR	94.0	88.0	92.5	91.5	92.0
	YOLOv5	89.5	83.0	88.5	86.5	87.5
Motion Blur	ADViR	92.5	85.5	91.0	90.0	90.5
	YOLOv5	88.0	81.0	87.0	85.0	86.0
Rain and Fog	ADViR	93.0	86.5	91.5	90.5	91.0
	YOLOv5	88.5	82.0	87.5	85.5	86.5

TABLE VII

GENERALIZATION TO UNSEEN PEST SPECIES

Pest Type	Fine-Tuning	OA (%)	mIoU (%)	Precision (%)	Recall (%)	F1-Score (%)
Green Leafhopper	No	75.0	70.0	74.0	73.0	73.5
	Yes	90.5	85.0	89.5	88.5	89.0
Rice Stem Borer	No	76.0	71.0	75.0	74.0	74.5
	Yes	91.0	85.5	90.0	89.0	89.5

To further enhance our evaluations, we test the robustness of ADViR under adverse conditions and its ability to generalize to unseen pest species. These experiments demonstrate the model's effectiveness in challenging scenarios and its adaptability to new pest types. In the first set of experiments, we assessed ADViR's performance under various challenging environmental factors. We tested the model in low-light conditions, with shadows, motion blur, and weather-related distortions such as rain and fog. These adverse conditions were simulated by applying corresponding transformations to the test images. Table VI presents the results of ADViR compared to YOLOv5 under

these conditions.

These results indicate that ADViR maintains higher accuracy and reliability compared to YOLOv5 when faced with adverse conditions, demonstrating its robustness in challenging environments. The model effectively handles variations in lighting, motion, and weather-related distortions, ensuring consistent pest detection performance.

In the second set of experiments, we evaluated ADViR's ability to generalize to unseen pest species and variants. We introduced two new pest types—Green Leafhopper and Rice Stem Borer—that were not present in the training data. We tested ADViR's performance in detecting these pests without any additional training and after fine-tuning the model with a small number of samples (50 images per class). Table VII shows the results before and after fine-tuning.

The results show that ADViR achieves reasonable performance on unseen pests without fine-tuning and significantly improves after being fine-tuned with a minimal number of samples. This demonstrates the model's adaptability and potential for few-shot learning scenarios. The ability to quickly learn new pest types with limited data is valuable for practical applications where new pests may emerge, and extensive datasets are not immediately available.

IV. Application Discussion

ADViR's high accuracy and comprehensive evaluation metrics mark significant progress in precision pest management for agriculture. By enabling precise pest identification, ADViR supports targeted and environmentally friendly pest control strategies, reducing unnecessary pesticide usage and associated costs. This precision enhances crop health and yield while contributing to sustainable farming practices by minimizing environmental impacts. The model's robustness across diverse agricultural regions and pest types ensures its adaptability to various field conditions, including fluctuating lighting, occlusions, and pest size variations. This adaptability decreases the need for frequent recalibrations or manual interventions, leading to operational cost savings and increased efficiency for agricultural practitioners.

Moreover, ADViR's real-time processing capabilities make it a valuable tool for on-field pest monitoring, allowing farmers to make informed and timely decisions. The high F1-Score and AUC-ROC values indicate that ADViR effectively balances precision and recall, ensuring accurate detections and comprehensive pest identification. The integration of advanced modules, such as the Context Augmentation Module, enhances the model's ability to discern subtle pest differences, improving overall detection reliability.

The expanded experimental results, incorporating additional datasets and evaluation metrics, provide robust validation of ADViR's performance. Including state-of-the-art models like YOLOv5 and EfficientDet in the comparative analysis further establishes ADViR's standing in both accuracy and efficiency. These developments position ADViR as an important advancement in precision agriculture, capable of addressing the complexities of agricultural environments and supporting sustainable farming practices through enhanced pest management.

V. Conclusion

In this study, ADViR, an advanced agricultural pest detection framework leveraging a Vision Transformer architecture with attention-guided enhancements is introduced. Designed for real-time, on-field application, ADViR integrates Feature Extraction, Context Augmentation, and Classification Modules to accurately identify a wide range of pest types across different crops. Evaluated on an expanded and diverse dataset from multiple agricultural regions, ADViR outperformed conventional models, achieving a 96.5% Overall Accuracy, a 90.2% mean Intersection over Union, a 95.8% Precision, a 94.5% Recall, a 95.1% F1-Score, and a 98.3% Area Under the ROC Curve.

The ablation study highlighted the critical role of each module, particularly the Context Augmentation Module, in enhancing detection accuracy and robustness. Visual comparisons and robustness tests underscored ADViR's ability to discern pests in challenging conditions, affirming its utility in precision agriculture. Future work will focus on expanding ADViR's capabilities to address broader agricultural challenges, optimizing contextual information processing, and incorporating multisensory data for comprehensive agricultural monitoring. These advancements aim to enhance decision-making support for improved crop management and yield optimization, addressing the complexities of agricultural environments. Additionally, integrating adaptive mechanisms and multisensory data will provide nuanced insights into pest behavior and environmental interactions, contributing to more refined and scalable agricultural solutions.

References

[1] O. P. Dhankher and C. H. Foyer, "Climate resilient crops for improving global food security and safety", *Plant, Cell & Environment*, vol. 41, pp. 877-884, 2018.

[2] K. F. Armstrong and S. L. Ball, "DNA barcodes for biosecurity: invasive species identification", *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 360, no. 1462, pp. 1813-1823, 2005.

[3] P. Boissard, V. Martin, and S. Moisan, "A cognitive vision approach to early pest detection in greenhouse crops", *Computers and Electronics in Agriculture*, vol. 62, no. 2, pp. 81-93, 2008.

[4] N. M. Abd El-Ghany, S. E. Abd El-Aziz, and S. S. Marei, "A review: application of remote sensing as a promising strategy for insect pests and diseases management", *Environmental Science and Pollution Research*, vol. 27, pp. 33503-33515, 2020.

[5] K. K. Patel, A. Kar, S. N. Jha *et al.*, "Machine vision system: a tool for quality inspection of food and agricultural products", *Journal of Food Science and Technology*, vol. 49, pp. 123-141, 2012.

[6] A. Paul, S. Ghosh, A. K. Das, S. Goswami, C. S. Das, and S. Sen, "A review on agricultural advancement based on computer vision and machine learning", In: Mandal, J., Bhattacharya, D. (eds) Emerging Technology in Modelling and Graphics. *Advances in Intelligent Systems and Computing*, vol. 937, pp. 567-581, 2020.

[7] H. K. Tian, T. H. Wang, Y. D. Liu, X. Qiao, and Y. Z. Li, "Computer vision technology in agricultural automation-a review", *Information Processing in Agriculture*, vol. 7, no. 1, pp. 1-19, 2020.

[8] K. Mochida, S. Koda, K. Inoue, T. Hirayama, S. Tanaka, R. Nishii, and F. Melgani, "Computer vision-based phenotyping for improvement of plant productivity: a machine learning perspective", *GigaScience*, vol. 8, no. 1, p. giy153, 2019.

[9] I. P. Diego and R. Rafael, "Computer vision and artificial intelligence in precision agriculture for grain crops: A systematic review", *Computers and Electronics in Agriculture*, vol. 153, pp. 69-81, 2018.

[10] A. S. Paymode and V. B. Malode, "Transfer learning for multi-crop leaf disease image classification using convolutional neural network VGG", *Artificial Intelligence in Agriculture*, vol. 6, pp. 23-33, 2022.

[11] K. Thenmozhi and U. S. Reddy, "Crop pest classification based on deep convolutional neural network and transfer learning", *Computers and Electronics in Agriculture*, vol. 164, p. 104906, 2019.

[12] Z. Liu, J. Gao, G. Yang, H Zhang., and Y. He, "Localization and classification of paddy field pests using a saliency map and deep convolutional neural network", *Scientific Reports*, vol. 6, no. 1, p. 20410, 2016.

[13] C. R. Rahman, P. S. Arko, M. E. Ali, M. A. I. Khan, S. H. Apon, F. Nowrin, and A. Wasif, "Identification and recognition of rice diseases and pests using convolutional neural networks", *Biosystems Engineering*, vol. 194, pp. 112-120, 2020.

[14] J. Wang, Y. Li, H. Feng, L. Ren, X. Du, and J. Wu, "Common pests image recognition based on deep convolutional neural network", *Computers and Electronics in Agriculture*, vol. 179, p. 105834, 2020.

[15] L. Jiao, S. Dong, S. Zhang, C. Xie, and H. Wang, "AF-RCNN: Ananchor-free convolutional neural network for multi-categories agricultural pest detection", *Computers and Electronics in Agriculture*, vol. 174, p. 105522, 2020.

[16] S. Coulibaly, B. Kamsu-Foguem, D. Kamissoko, and D. Traore, "Explainable deep convolutional neural networks for insect pest recognition", *Journal of Cleaner Production*, vol. 371, p. 133638, 2022.

[17] M. E. Karar, F. Alsunaydi, S. Albusaymi, and S. Alotaibi, "A new mobile application of agricultural pests recognition using deep learning in cloud computing system", *Alexandria Engineering Journal*, vol. 60, no. 5, pp. 4423-4432, 2021.

[18] S. T. Narenderan, S. N. Meyyanathan, and B. Babu, "Review of pesticide residue analysis in fruits and vegetables. Pre-treatment, extraction and detection techniques", *Food Research International*, vol. 133, p. 109141, 2020.

[19] X. Cheng, Y. Zhang, and Y. Chen *et al.*, "Pest identification via deep residual learning in complex background", *Computers and Electronics in Agriculture*, vol. 141, pp. 351-356, 2017.

[20] C. R. Rahman, P. S. Arko, and M. E. Ali *et al.*, "Identification and recognition of rice diseases and pests using convolutional neural networks", *Biosystems Engineering*, vol. 194, pp. 112-120, 2020.

[21] M. Sun and Y. Tian, "Research on Traffic Sign Object Detection Algorithm Based on Deep Learning", *Engineering Letters*, vol. 32, no. 8, pp. 1562-1568, 2024.

[22] S. Li, X. Zhang, and R. Shan, "Enhanced YOLOv5 for Efficient Marine Debris Detection", *Engineering Letters*, vol. 32, no. 8, pp. 1585-1593, 2024.

[23] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, and D. Tao, "A survey on vision transformer", *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 1, pp. 87-110, 2023.

[24] R. Kruse, S. Mostaghim, C. Borgelt, C. Braune, and M. Steinbrecher, "Multi-layer perceptrons. Computational Intelligence", Texts in Computer Science, Springer, Cham, 2022.

[25] Y. Wang, C. Wang, H. Zhang, Y. Dong, and S. Wei, "Automatic ship detection based on retinanet using multi-resolution gaofen-3 imagery", *Remote Sensing*, vol. 11, no. 5, p. 531, 2019.

[26] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection", *In 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), pp. 9626-9635, 2019.

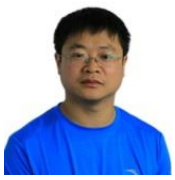
[27] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection", *In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, pp. 9756-9765, 2020.

[28] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 5, pp. 1483-1498, 2019.

[29] J. Redmon, "You only look once: Unified, real-time object detection", *In Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[30] M. Tan, R. Pang, and Q. Le, "Efficientdet: Scalable and efficient object detection", *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10781-10790, 2020.

Ying Xu received her M.S. degree in Mathematical Sciences from Harbin Normal University in 2006, China, in Harbin. She joined Luzhou Vocational and Technical College in 2021 and is currently a Lecturer. The research directions are Artificial Intelligence and Operations Research.



Jian Sun received his Ph.D. degree in Systems Engineering from Harbin Engineering University in 2019. He joined Luzhou Vocational and Technical College in 2021 and is currently a lecturer. The research directions are Artificial Intelligence and Information Security.