



REVIEW

# Intelligent Crop Pest Identification: From Traditional Machine Learning to Multimodal Large Models—A Review

Shuai Zhou<sup>1,2,3</sup>, Jun Wu<sup>3,4</sup>, Li Li<sup>2,5</sup>, Rong Tang<sup>1,2</sup>, Zhangjun Peng<sup>1,2</sup>, Mingfei Wan<sup>1,2,3</sup>,  
Zhiqiang Chen<sup>5</sup> and Zhigui Liu<sup>1,2,\*</sup>

<sup>1</sup>College of Information and Control Engineering, Southwest University of Science and Technology, Mianyang, China

<sup>2</sup>Sichuan Engineering Technology Research Center of Industrial Self-Supporting and Artificial Intelligence, Mianyang, China

<sup>3</sup>Mianyang Zhongke Huinong Digital Intelligence Technology Co., Ltd., Mianyang, China

<sup>4</sup>College of Life Sciences and Agri-Forestry, Southwest University of Science and Technology, Mianyang, China

<sup>5</sup>College of Computer Science and Technology, Southwest University of Science and Technology, Mianyang, China

\*Corresponding Author: Zhigui Liu. Email: liuzhigui@swust.edu.cn

Received: 16 June 2026; Accepted: 31 August 2026; Published: 22 October 2026

**ABSTRACT:** Crop pests pose a major threat to agricultural production and hinder the sustainable development of agriculture. Accurate pest identification is a prerequisite for implementing sustainable pest control strategies and enabling precise and efficient pest management. This review examines the technological evolution of intelligent crop pest identification and analyzes the advantages and limitations of existing methods. First, it analyzes three types of crop pest data collection methods, with emphasis on sampling bias, existing bottlenecks such as long-tail distribution and domain shift, and the characteristics of typical datasets. Second, it discusses the application scenarios and limitations of traditional machine learning methods, including clustering, support vector machines, and decision trees. Next, it compares the technical characteristics and performance differences of one-stage and two-stage convolutional neural networks, as well as Vision Transformer models, in pest identification. Then, it discusses the decoupled and end-to-end architectures of multimodal large models and their cross-modal applications in intelligent pest diagnosis, thereby offering new perspectives for closed-loop pest diagnosis and ecologically based prevention and control decision-making. Finally, this review summarizes the limitations of existing research and proposes future research directions for developing more reliable and efficient crop pest identification technologies.

**KEYWORDS:** Pest identification; deep learning; machine learning; multimodal large models

## 1 Introduction

Agriculture serves as a core cornerstone of global economic and social development, and the stability and high quality of its output are paramount for maintaining supply chain security and driving sustainable development. However, in recent decades, climate change and ecological environment alterations have intensified the severity of crop pests, posing a severe threat to global food security. According to monitoring data from the National Crop Pest Forecasting Network, the total occurrence area of pests and diseases on major food crops (wheat, rice, corn, potato) in China is projected to reach 2.518 billion mu in 2025, an increase of 6.2% compared to 2024, with over 70% of grain-producing areas at risk of infestation [1]. Globally, the Food and Agriculture Organization (FAO) estimates that pests cause annual crop losses of 15%–20% of total production, resulting in direct economic losses exceeding hundreds of billions of US dollars—with even higher loss ratios in developing countries due to backward prevention and control technologies [2].

Traditional pest identification and control models have significant limitations and cannot meet the needs of modern agricultural development. For decades, pest identification has relied on manual visual inspection and empirical judgment by agricultural technicians or farmers. This approach is time-consuming and labor-intensive, and its accuracy is highly dependent on subjective experience and professional expertise—especially for morphologically similar related species or larval-stage pests, which are prone to misjudgment [3]. Additionally, manual identification has a limited coverage range, making it difficult to achieve large-area, high-frequency monitoring; this often results in missed optimal control opportunities during the early stages of pest outbreaks [4]. Although some regions have introduced traditional machine learning-based identification technologies, these methods rely on manually designed features and exhibit poor adaptability to background interference, pest posture changes, and uneven lighting in complex field environments, failing to meet the accuracy and robustness requirements of practical production [5,6].

The rapid development of artificial intelligence technology has provided a new technological paradigm for solving the problem of crop pest identification. Deep convolutional neural networks (DCNN) have achieved breakthrough progress in the field of computer vision with their powerful automatic feature extraction and complex pattern recognition capabilities. Unlike traditional approaches, DCNNs can automatically extract deep semantic features from images through multiple layers of nonlinear transformations without the need for manual intervention, and can achieve mapping from pixel-level to semantic-level information. For example, in natural image recognition, models based on DCNN have achieved recognition accuracy exceeding human level on large datasets such as ImageNet, laying a solid foundation for their application in agricultural scenarios [6,7].

The development of intelligent crop pest identification has experienced several technological transitions, as illustrated in Fig. 1. Early studies mainly relied on traditional machine learning methods, where hand-crafted features combined with classical classifiers were used for pest recognition. Although these approaches achieved acceptable performance under controlled conditions, their dependence on manually designed features limited their adaptability to complex agricultural environments. With the rapid development of deep learning after 2012, convolutional neural networks gradually replaced traditional feature engineering by enabling automatic feature extraction and more robust visual representation learning. Subsequently, object detection frameworks such as Faster R-CNN and YOLO further extended pest identification from image-level classification to precise localization of pest targets.

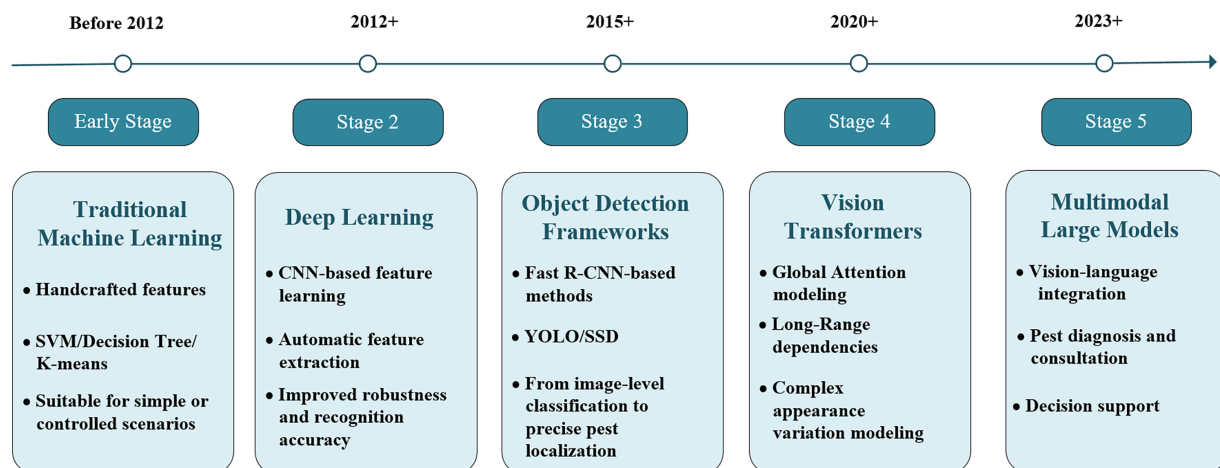


**Figure 1:** Evolution of intelligent crop pest identification technologies.

Since 2020, Vision Transformer-based methods have introduced global attention mechanisms into visual recognition tasks, providing new solutions for modeling complex pest appearance variations and long-range dependencies. More recently, multimodal large models have emerged as a promising direction by integrating visual perception with language-based knowledge reasoning, enabling not only pest recognition but also intelligent diagnosis and decision support. This evolution reflects the transition of pest identification technologies from handcrafted feature extraction toward automated perception, multimodal understanding, and intelligent agricultural decision support.

The rapid development of agricultural informatization, IoT devices, UAVs, high-definition imaging, edge computing, and mobile chips has provided both data and hardware support for intelligent pest identification [8–10]. Previous reviews have mainly focused on pest detection algorithms and computer vision applications. In contrast, this review analyzes the field from three perspectives: data characteristics, algorithm evolution, and practical deployment. Previous studies have provided valuable discussions on pest detection algorithms and related challenges; however, the systematic integration of data-related issues, emerging intelligent models, and real-world deployment considerations remains limited. In this review, we systematically analyze the sampling biases introduced by different pest image acquisition strategies, including Internet-based collection, pest monitoring systems, and field image acquisition, and discuss their impacts on long-tail distribution, domain shift, and model generalization. Furthermore, this review extends the discussion beyond conventional visual recognition models by analyzing the evolution toward Vision Transformers and multimodal large models, highlighting the transition from image-based pest identification to multimodal agricultural diagnosis and closed-loop intelligent decision-making architectures. Finally, we summarize the challenges associated with real-world deployment, including complex field environments, computational constraints, and cross-domain generalization, providing future perspectives for developing reliable and intelligent pest identification systems.

## 2 Review Methodology

In this review, a structured literature collection and analysis strategy was adopted to comprehensively summarize the development of intelligent crop pest identification technologies. As illustrated in Fig. 2, the overall methodology consisted of four main steps: literature retrieval, manual search, literature screening, and classification analysis.

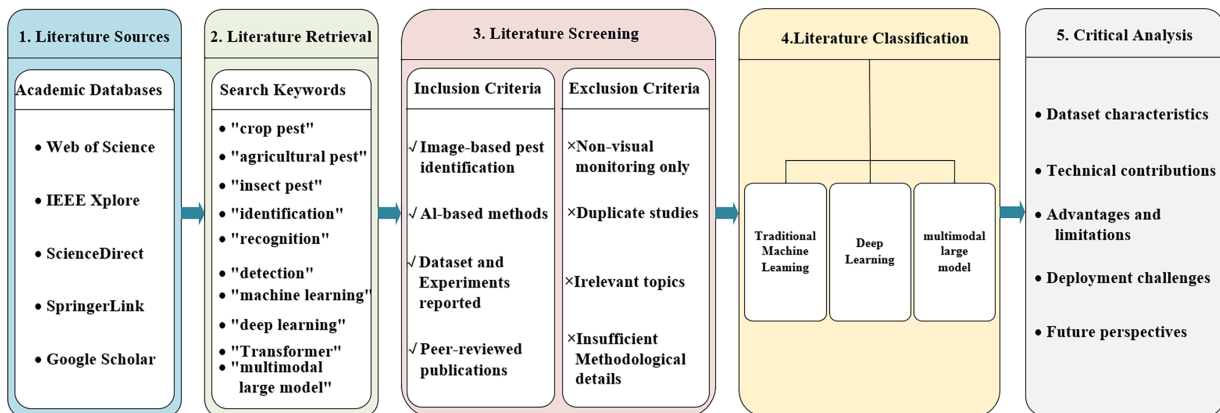


**Figure 2:** Overview of the literature search and analysis methodology.

## 2.1 Initial Literature Retrieval

Relevant publications were collected from multiple academic databases, including Web of Science Core Collection, IEEE Xplore, ScienceDirect, SpringerLink, and Google Scholar. The literature search mainly focused on studies published between 2010 and 2026, covering the development of traditional machine learning, deep learning, Vision Transformer, and multimodal large model-based pest identification technologies. The primary keywords included combinations of “crop pest”, “agricultural pest”, “insect pest”, “identification”, “recognition”, “detection”, “machine learning”, “deep learning”, “object detection”, “Transformer”, and “multimodal large model”.

## 2.2 Manual Literature Search

To avoid missing important studies that may not be retrieved through database searching, additional manual searches were conducted. Specifically, the references of highly cited papers, recent review articles, and representative studies were examined. In addition, relevant conference proceedings and journal publications were manually checked to identify important technological developments and representative algorithms in the field of intelligent pest identification.

## 2.3 Literature Screening Process

The initial search resulted in a large number of candidate publications. These studies were first screened by examining titles and abstracts to remove irrelevant papers. Subsequently, the full texts of the remaining publications were carefully reviewed according to predefined inclusion and exclusion criteria.

The inclusion criteria were as follows:

- (1) The study focused on image-based crop pest identification, recognition, or detection tasks;
- (2) Artificial intelligence-based methods were proposed or evaluated;
- (3) Sufficient descriptions of datasets, algorithms, and experimental results were provided;
- (4) The study was published in a peer-reviewed journal or high-quality conference.

The exclusion criteria included:

- (1) Studies focusing only on non-visual pest monitoring without identification tasks;
- (2) Duplicate publications;
- (3) Studies lacking sufficient methodological details;
- (4) Research unrelated to intelligent crop pest identification.

## 2.4 Literature Classification and Analysis

The selected studies were categorized according to the evolution of intelligent pest identification technologies, including traditional machine learning methods, deep learning-based approaches, Vision Transformer-based methods, and multimodal large model-based solutions. Furthermore, the reviewed studies were analyzed from multiple perspectives, including dataset characteristics, technical contributions, application scenarios, advantages and limitations, and deployment challenges.

The overall literature collection and analysis framework is illustrated in [Fig. 2](#). This methodology enables a comprehensive and balanced understanding of the technological evolution, current challenges, and future development directions of intelligent crop pest identification.

### 3 Data Collection Methods for Crop Pests and Publicly Available Datasets

Deep learning has made remarkable progress in the fields of image recognition and object detection, largely benefiting from the support of large-scale annotated datasets. These datasets provide a unified evaluation platform for algorithm research, thereby enabling continuous innovation. Benchmark datasets such as ImageNet have led to qualitative breakthroughs in general visual tasks.

In the field of crop pest detection, however, the complex and variable farm environments, as well as challenges such as fluctuating illumination and cluttered backgrounds, make data collection and annotation considerably more difficult. Compared with general visual datasets, pest images often feature small targets, dense distributions, and severe occlusions, which impose higher requirements on data acquisition and sample quality. To obtain image samples suitable for different pest identification tasks, researchers have developed multiple data collection strategies, including automated monitoring systems, field manual acquisition, and web-based image collection.

Currently, public pest datasets are primarily constructed utilizing the following methodologies:

- (1) Pest monitoring lights. Continuous monitoring is conducted at specific observation points using monitoring lights, and pests are automatically captured through light attraction. This method enables long-term data collection, but it is usually point-based and has limited spatial coverage. Most of the collected samples are adult insects.
- (2) Manual collection. Pest image samples are collected in the production fields using digital cameras or other mobile shooting devices. The manual collection method can obtain real field pest image samples, but it requires substantial labor and has relatively low efficiency.
- (3) Internet collection. Pest images are obtained from the Internet (including agricultural research institutions, open-source platforms or social media), and then manually screened and labeled to build the dataset. Network-based collection can rapidly obtain large volumes of pest images. However, because these images come from heterogeneous sources, their quality varies considerably, making it difficult to meet the requirements of real-field applications.

The visual characteristics of pest images collected using these three approaches differ substantially because they are designed for different application scenarios. As shown in Fig. 3a, web-collected images usually contain clear and representative pest instances with relatively simple backgrounds, which are valuable for species classification and large-scale dataset construction. However, their similarity to real-field environments depends on the specific research objectives. In contrast, images obtained by pest monitoring lamps (Fig. 3b) are typically collected under automated and fixed monitoring conditions, often containing dense targets, occlusion, and debris interference. Such characteristics are beneficial for automated pest surveillance systems but require strong small-object detection capability. Manual field acquisition (Fig. 3c) provides more realistic agricultural scenarios with complex backgrounds and multi-scale variations, making it suitable for evaluating model robustness in practical environments. These differences among data sources represent distinct domain characteristics, which may influence model generalization when models are transferred across different application scenarios.

The three pest data collection approaches described above exhibit different sampling characteristics due to their acquisition principles, sampling conditions, and intended applications. These characteristics lead to variations in dataset distributions, which should be considered when designing and evaluating recognition models. Importantly, such differences should not be regarded as inherent limitations of the collection methods; instead, they reflect the specific purposes and operating conditions of each acquisition strategy. For pest monitoring lamps, the sampling characteristics are mainly reflected in species and temporal selectivity. Such equipment primarily captures nocturnal phototactic pests, whereas diurnal or non-phototactic pests are

less represented. In addition, because monitoring lamps are usually deployed at fixed observation locations and operated during specific monitoring periods, the collected images mainly reflect pest populations under these conditions. Therefore, models trained on monitoring-lamp datasets may require further adaptation when applied to different pest communities or broader agricultural environments.

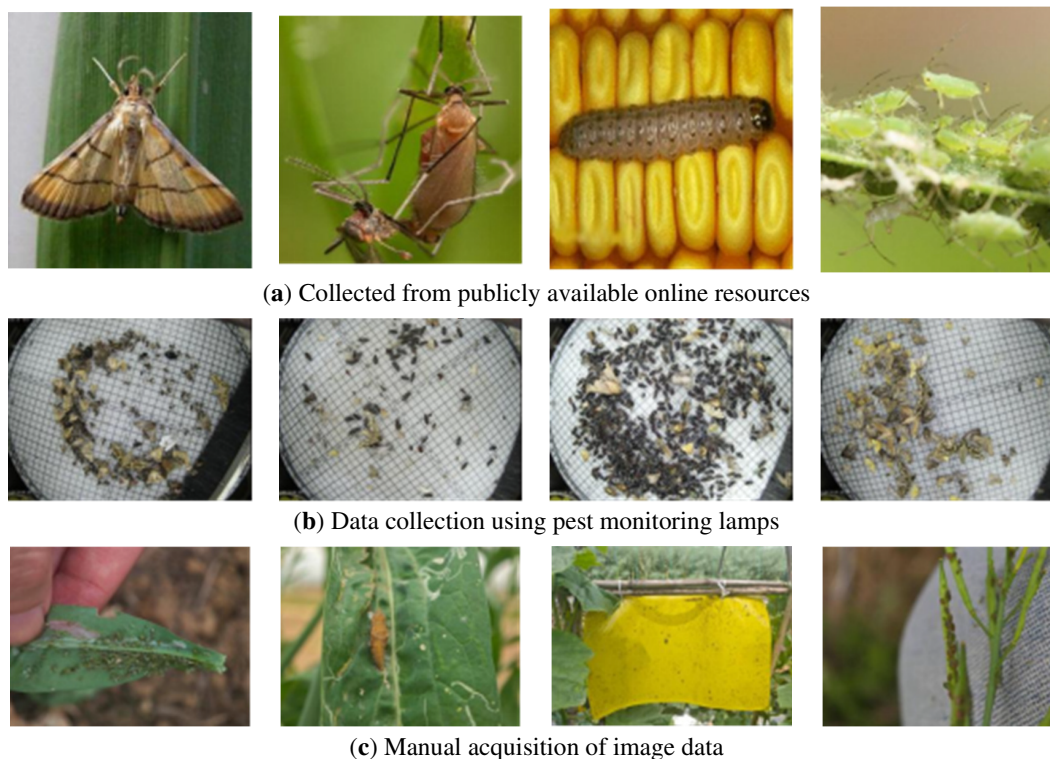


**Figure 3:** Three different methods of collecting pest data.

For manual field collection, dataset characteristics are mainly influenced by geographic location, seasonal conditions, and collection strategies. Images collected in specific crop fields and growth periods can provide realistic representations of agricultural environments, including natural backgrounds, different pest postures, and scale variations. However, such datasets may also contain regional and seasonal characteristics. When models are applied to other regions, crop varieties, or growth stages, distribution differences between training and deployment environments may affect recognition performance.

For web-based collection, the main characteristics originate from heterogeneous image sources and acquisition conditions. Images collected from agricultural databases, research publications, and online platforms often emphasize clear visualization of pest morphology, which facilitates species recognition and increases category diversity. Nevertheless, these images may differ from field scenarios involving complex backgrounds, occlusion, and variable illumination. Therefore, models trained mainly on web-collected datasets should be carefully evaluated when deployed in real agricultural environments.

Overall, each data acquisition strategy has its own advantages and application value. Automated monitoring systems provide continuous observations for pest surveillance, manual field collection captures realistic agricultural environments, and web-based collection improves species coverage and dataset accessibility. Therefore, the selection of data collection methods should be determined according to the target application scenario. Furthermore, combining multiple acquisition strategies can provide complementary information and improve the generalization capability of intelligent pest identification models.

Based on sample size, species coverage, open access and scenario characteristics summarized in Table 1, the representative datasets for crop pest identification can be categorized into three types. The first category comprises general-purpose open datasets with high usability, represented by AP162 and IP102. Characterized by large sample sizes, rich species coverage and open access, these datasets serve as high-quality benchmark datasets for pest identification research, with only a slight bias of long-tailed sample distribution. The second type refers to scenario-specific open datasets with targeted usability, typified by AgriPest, Pest24 and other relevant datasets. Endowed with clear scenario positioning and detailed sample annotations, they are suitable for research on specific pest identification tasks, yet suffer from such limitations as narrow species coverage and small sample sizes. The third type is closed datasets with limited usability and low reliability, represented by MPD2018, LMPD2020 and others. Plagued by drawbacks including non-open access and small sample sizes, these datasets have relatively low practical application value and are only applicable to internal research within their developing institutions. In addition, the Persimmon Dataset released in 2020 is a typical small-scale open dataset for crop-specific pest identification, which can provide data support for research on pest recognition of characteristic cash crops. Although remarkable progress has been made in the development of crop pest datasets in terms of sample size and species coverage, the field still faces common bottlenecks such as long-tailed sample distribution, insufficient samples from complex field scenarios, and the lack of samples of rare and regional pest species. For the future development of such datasets, efforts should be centered on the actual demands of agricultural production: the multi-source fusion collection method should be adopted as the primary approach to supplement the aforementioned deficient samples, and high-quality datasets should be openly shared, thereby laying a solid data foundation for the practical application of intelligent pest identification technologies.

**Table 1:** Overview of existing public and private image datasets for crop pest recognition.

| Year | Dataset                | Scale ( $\times 10^3$ Images) | Number of Classes | Collection Method     | Task Type                           | Open |
|------|------------------------|-------------------------------|-------------------|-----------------------|-------------------------------------|------|
| 2018 | MPD2018 [11]           | 88.6                          | 16                | Pest monitoring lamps | Object detection and classification | No   |
| 2021 | AgriPest [12]          | 49.7                          | 14                | Field acquisition     | Detection                           | Yes  |
| 2019 | IP102 [13]             | 18.9                          | 102               | Internet collection   | Object detection and classification | Yes  |
| 2022 | IP67 [14]              | 6.8                           | 67                | Internet collection   | Classification                      | No   |
| 2018 | D0 [15]                | 4.5                           | 40                | Field acquisition     | Classification                      | Yes  |
| 2020 | LMPD2020 [16]          | 1.8                           | 26                | Pest monitoring lamps | Detection                           | No   |
| 2020 | Persimmon Dataset [17] | 0.6                           | 8                 | Internet collection   | Classification                      | Yes  |
| 2020 | Pest24 [18]            | 25.4                          | 24                | Pest monitoring lamps | Detection                           | Yes  |
| 2021 | TPest [19]             | 1.4                           | 2                 | Field acquisition     | Detection                           | Yes  |
| 2021 | Rustia2021 [20]        | 0.9                           | 4                 | Pest monitoring lamps | detection                           | Yes  |
| 2025 | AP162 [21]             | 194                           | 162               | Internet collection   | Classification                      | Yes  |

In natural environments, the occurrence frequency of pests and the quantity of collectable samples vary significantly due to factors such as geographical distribution, seasonal cycles, host crop species and the focus of artificial monitoring. This inherent disparity causes an imbalance in dataset distribution at the source, which is an objective reflection of the real-world condition that some pests are common dominant species while most are regional or occasional species, thus forming a typical long-tail phenomenon where a small number of categories account for a large number of samples and the vast majority of categories suffer from scarce samples. As visualized in Fig. 4, this phenomenon is particularly prominent in the two representative datasets IP102 and Pest24: the sample distribution chart of IP102 clearly shows that the sample size of a few head categories of pests is close to 2000, while that of the vast majority of long-tail categories is less than 500 or even nearly zero, and the category name distribution of Pest24 in the figure also intuitively reflects its potential long-tail distribution characteristics. Such severe data imbalance poses great challenges to model training, which not only makes it difficult for the model to effectively identify pest categories with scarce samples, but also significantly reduces its recognition performance for rare or emerging pest species. Absent deliberate action such as class-balanced loss functions and few-shot adaptation mechanisms, models being trained using these biased datasets will have difficulties in attaining high levels of generalization performance in the dynamic real-life agricultural scenarios.

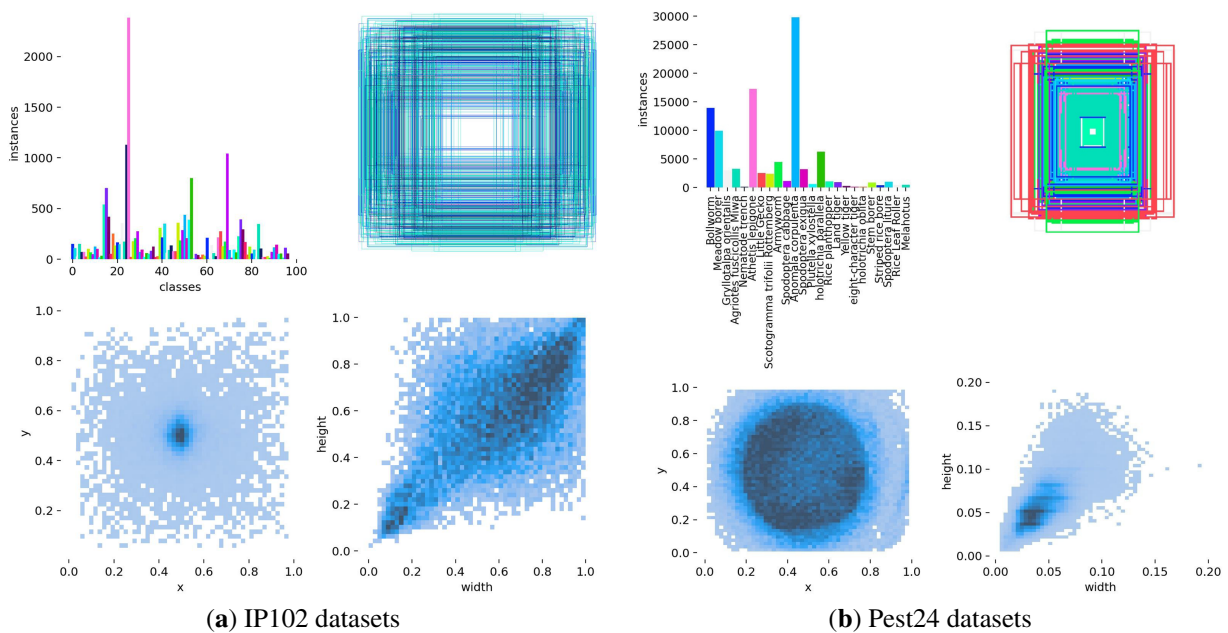


**Figure 4:** Visualization of the analysis of the IP102 and Pest24 datasets.

#### 4 Traditional Machine Learning Approaches

Traditional machine learning represents the earliest core technical approach in the field of automatic crop pest identification. Relying on handcrafted image features and classical statistical classifiers, such methods fulfill the tasks of pest image recognition and counting. While featuring low computational cost and stable performance on small-sample datasets, they are highly dependent on manual feature engineering and fail to maintain satisfactory robustness in complex field scenarios with uneven illumination, cluttered backgrounds, variable pest postures and severe mutual occlusion. This section focuses on three mainstream traditional machine learning methods, namely clustering, support vector machines and decision trees, to analyze their practical performance and applicable boundaries in pest identification tasks.

#### 4.1 Clustering-Based Methods

Clustering-based crop pest detection methods first conduct statistical analysis on the size, shape or feature distribution of pest targets in training samples, and generate representative target categories or prior parameters through clustering algorithms, as illustrated in Fig. 5. The obtained findings of the clustering are subsequently incorporated into feature learning and prediction processes of the detection network, which are used to direct multi-scale feature matching and localization of targets, and eventually, achieve the detection and recognition of pest targets.

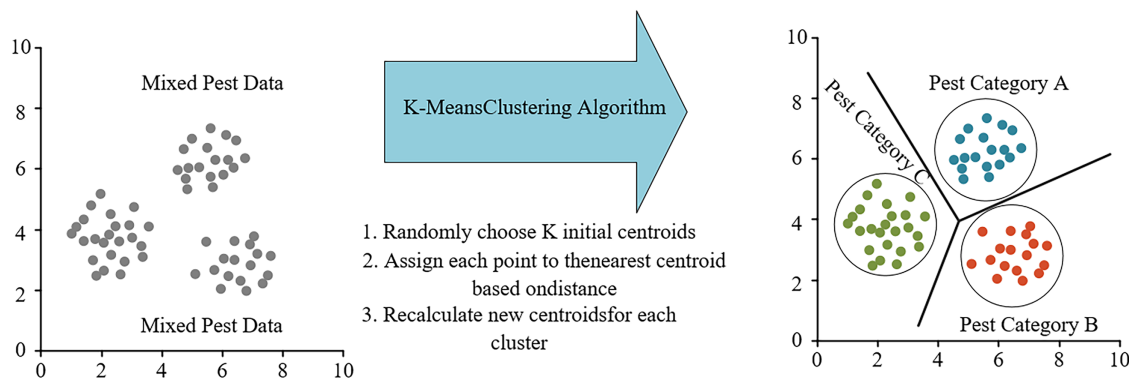


**Figure 5:** Schematic diagram of pest data clustering using the K-means algorithm.

Relying on the above core principles, clustering algorithms have been applied in a number of practical tasks for pest identification and monitoring. For example, Watts and Worner [22] optimized the insect invasion risk assessment model using K-means clustering and self-organizing maps, and took the species frequency in clusters as the risk weight to effectively improve the assessment accuracy. Based on the idea of clustering-driven feature analysis, Li et al. [23] further established a K-means clustering recognition pipeline for citrus red spiders (*Panonychus citri*): by converting RGB images to the Lab color space and performing clustering analysis on the a and b channels, this method demonstrated effective recognition performance on high-definition images with 5 clustering centers.

Extending from qualitative identification to quantitative pest monitoring, clustering technology has also become a reliable tool for pest counting. In typical scenarios, Wang et al. [24] applied K-means clustering to whitefly image segmentation, and then extracted the contours of whitefly clusters and estimated the number of individuals through ellipse fitting, which greatly improved the efficiency of traditional manual counting. To support field population monitoring, Liu et al. [25] also combined K-means clustering with global threshold segmentation to realize automatic counting of spider mites from field-collected images, providing a feasible solution for rapid field investigation.

Despite the phased application achievements of clustering methods in the fields of risk assessment, species identification and pest counting, they still have unavoidable inherent limitations. Such methods are highly sensitive to the selection of color space and the setting of clustering center parameters, and are easily disturbed by complex backgrounds such as uneven illumination, leaf veins and soil particles. In addition, clustering methods are difficult to adapt to variable pest postures and mutual occlusion between individuals, which greatly limits their practical application in complex agricultural environments [26].

With the rapid development of deep learning technologies represented by neural networks, their powerful end-to-end automatic feature extraction capability quickly established a performance advantage in complex scenarios. As a result, the use of clustering algorithms as standalone approaches in pest detection

research appears to have become less common. At present, these algorithms are increasingly integrated into modern detection frameworks as auxiliary tools, although they may still serve as core methods in certain specialized applications.

#### 4.2 SVM-Based Methods

Support Vector Machines (SVM) perform classification by finding the optimal hyperplane that maximizes the margin between classes. Owing to their favorable generalization performance on small-scale datasets, SVMs have been widely adopted in pest recognition tasks.

For remote monitoring requirements, Han [27] developed an SVM-based pest detection system for the Internet of Things (IoT). This system directly models raw image data without manual feature extraction, and realizes remote pest detection combined with image segmentation. For thrips monitoring in strawberry greenhouses, Ebrahimi et al. [28] further optimized the kernel function of SVM, extracted pest regions through image segmentation, and achieved a classification accuracy of over 90%.

For fine-grained pest identification tasks demanding higher discriminative capability, Yang [29] fused morphological features (invariant moments), color features (HSV histograms) and texture features (gray-level co-occurrence matrices) to construct an SVM classifier, which attained an accuracy of 88.3% in identifying early instar nymphs of *Oxya chinensis* and other overwintering pests. On this basis, Barman et al. [30] applied SVM to classify the fused features (gray-level co-occurrence matrices + gray-level run-length matrices) of coconut pest images, achieving nearly 100% recognition accuracy for five common coconut pest species.

Overall, although SVM-based methods can meet the basic needs of pest identification in small-scale and simple scenarios, they still have non-negligible application bottlenecks. Excessive reliance on handcrafted features makes it difficult for such methods to adapt to the complex morphological variations and high similarity of pest species in field environments. Meanwhile, the low computational efficiency of SVMs also restricts their extension to large-scale, high-throughput pest monitoring tasks, reducing the practical value of these methods in real agricultural production [31]. Nevertheless, relevant studies on SVMs have still been reported in recent years, yet SVMs are no longer constructed as independent pest identification frameworks, but serve as auxiliary classification modules integrated with emerging cutting-edge technologies (Fig. 6). In 2024, Chen et al. [32] proposed a soybean seed pest detection method based on spatial frequency domain imaging combined with reinforcement learning-optimized SVM (RL-SVM). This method used a spatial frequency domain imaging system to obtain subcutaneous images, optical absorption coefficient and reduced scattering coefficient of soybean seeds, and automatically optimized the parameters of the SVM classifier through a reinforcement learning algorithm, realizing high-precision graded detection of soybean seed pests with different damage degrees caused by *Riptortus pedestris* (especially concealed injuries that are difficult to distinguish by appearance). In 2025, Shen et al. [33] proposed a blockchain-assisted federated learning-driven SVM framework (BAFL-SVM) for smart agriculture. This framework extracted feature images via mean-shift image segmentation, built a classification model by integrating ResNet18 and SVM, and ensured secure data sharing on the blockchain through homomorphic encryption and secret sharing schemes, realizing efficient detection of crop pests on real rice pest datasets.

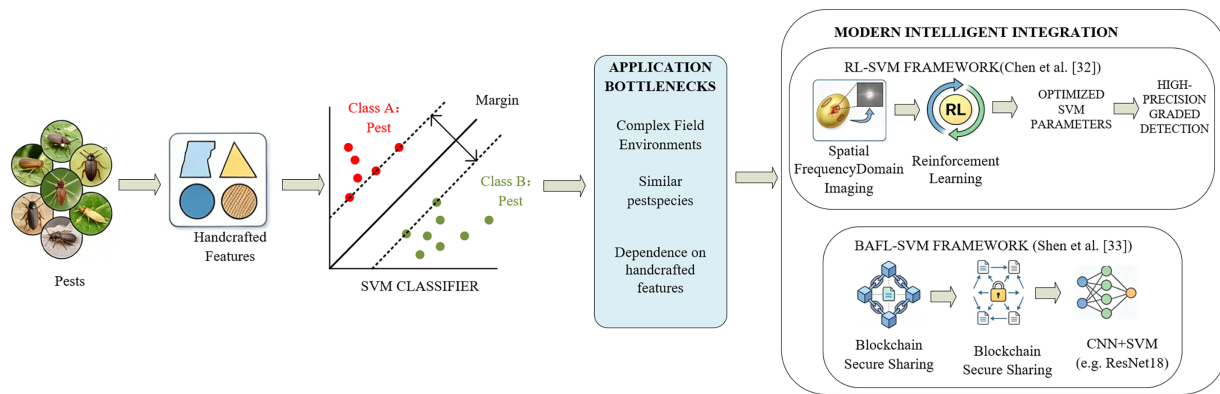


**Figure 6:** Schematic diagram of the evolution and optimization of SVM-based crop pest recognition methods [32,33].

### 4.3 Decision Tree-Based Methods

Decision trees establish classification rules in a tree structure and are widely used in pest counting and monitoring, facility pest detection, and field damage assessment due to their strong interpretability.

Ye [34] proposed a Vegetable Pest Counting Algorithm based on Machine Learning (VPCA-ML) for the needs of vegetable pest monitoring and counting. This algorithm integrates color features in the HSV color space with machine learning techniques such as decision trees to perform intelligent detection and pest counting on real-time field images of vegetable pests. In this algorithm framework, decision trees are used to construct classification models for distinguishing different pest species, realizing automatic identification and accurate counting of various pests on vegetable leaves in farmland.

For crop protection in greenhouse environments, Bu et al. [35] proposed a pest detection method for greenhouse tomatoes. This method first collects pest images with uniform backgrounds using insect sticky boards, then segments individual pest targets and extracts multi-dimensional image features including color, shape and texture. Finally, a random forest classification algorithm is adopted to identify pest species in image regions based on these features. As illustrated in Fig. 7, Yang et al. [36] further optimized the application form of decision trees by constructing an ensemble decision tree model based on multi-color spaces (HSI + Lab), combined with Prewitt and Canny operators for image segmentation, and finally achieved efficient classification of greenhouse pests with a recognition accuracy of 95.73%.

For field production scenarios with more complex background interference, Bütüner et al. [37] focused on the damage monitoring of *Tuta absoluta* and proposed a field damage intensity assessment method using the Decision Tree (DT) machine learning algorithm. This method directly trains and models complex environments containing backgrounds such as soil, weeds and stems using pixel matrix values of field images. Without manual background removal and handcrafted feature extraction, it achieves high-precision classification and recognition of leaf miner galleries on tomato leaves, providing an efficient solution for early pest monitoring in precision agriculture.

Decision trees and their derived methods play an important role in the prediction and identification of agricultural pests. By fusing multi-source data such as image visual features, hyperspectral remote sensing data and multi-color space information, combined with feature importance analysis and structured modeling capabilities, these methods realize the prediction of pest occurrence rules, pest classification and accurate target identification. Research focuses on the optimization of multi-dimensional feature extraction, multi-source data correlation modeling and the application of ensemble learning strategies. When used for pest detection, decision tree models are simple and interpretable, can process different types of data with high

computational efficiency, have low requirements for data standardization, can handle missing values, and are suitable for small datasets. However, they are sensitive to data noise and sample imbalance, which easily leads to model overfitting and unstable results, and their performance degrades as the input dimension increases, while their probability estimates may be unreliable, particularly when leaf nodes contain few samples or the data are imbalanced. Their performance is limited when dealing with high-dimensional or complex nonlinear relationships. The model complexity increases significantly with the growth of data scale, resulting in low deployment efficiency in field scenarios with high real-time requirements or dynamic changes, which limits their large-scale application potential [38].

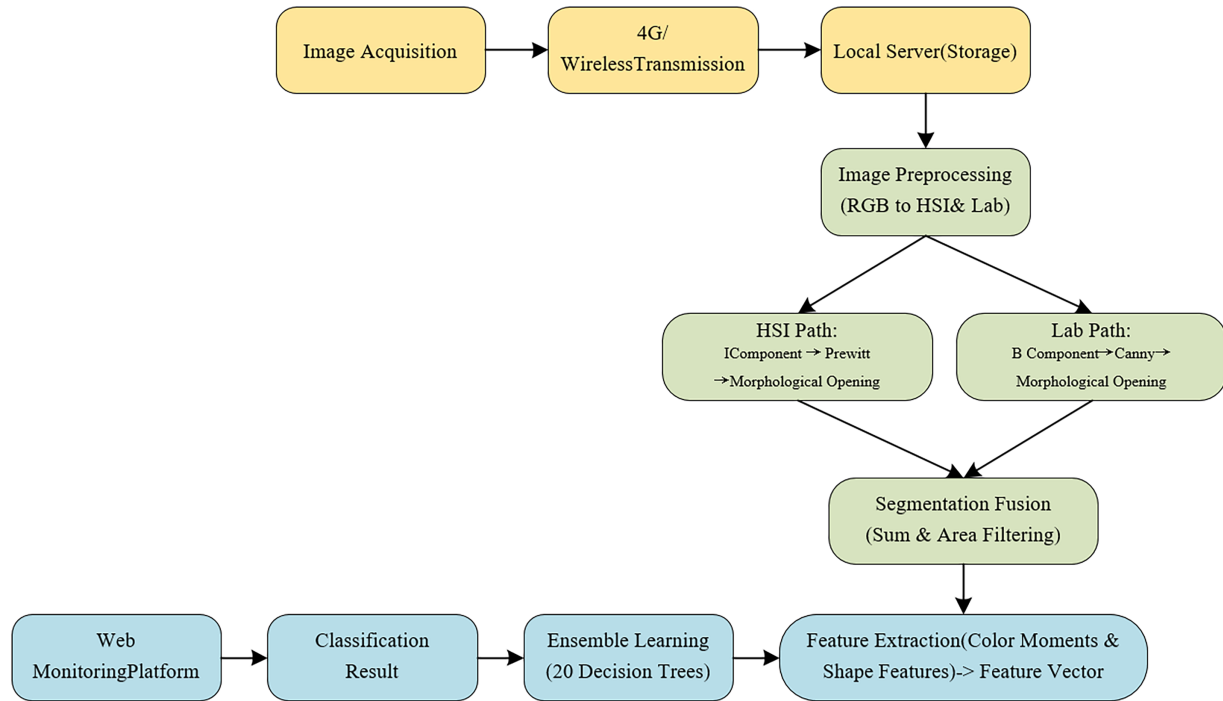


**Figure 7:** Greenhouse pest classification with dual-path preprocessing.

To provide a clearer comparison of traditional machine learning approaches, Table 2 summarizes the main applications, advantages, limitations, and representative studies of clustering-based methods, support vector machines, and decision trees in crop pest identification.

**Table 2:** Comparison of traditional machine learning methods for crop pest identification.

| Method                   | Main Applications                                           | Advantages                                                       | Limitations                                                                   | Representative Studies                                       |
|--------------------------|-------------------------------------------------------------|------------------------------------------------------------------|-------------------------------------------------------------------------------|--------------------------------------------------------------|
| Clustering-Based Methods | Pest segmentation, counting, and preliminary identification | Simple implementation; no requirement for extensive labeled data | Sensitive to feature selection, illumination changes, and complex backgrounds | K-means-based pest segmentation and counting methods [22–24] |

(Continued)

**Table 2 (continued)**

| Method                         | Main Applications                                                 | Advantages                                                                                    | Limitations                                                                                         | Representative Studies                                       |
|--------------------------------|-------------------------------------------------------------------|-----------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------|--------------------------------------------------------------|
| Support Vector Machines (SVMs) | Pest classification and recognition based on handcrafted features | Good generalization ability with small datasets; effective in high-dimensional feature spaces | Strong dependence on manually designed features; limited adaptability to complex field environments | SVM-based pest classification and monitoring methods [27–29] |
| Decision Trees (DT)            | Pest classification, counting, and decision support               | High interpretability; simple structure; easy feature integration                             | Prone to overfitting; sensitive to noisy data and class imbalance                                   | DT-based pest recognition and monitoring methods [34–37]     |

Overall, traditional machine learning methods established the foundation of intelligent pest identification by reducing dependence on manual inspection. However, their performance largely relies on handcrafted feature extraction and predefined statistical assumptions, which limits their robustness under complex agricultural environments. These limitations have motivated the transition toward deep learning methods with automatic feature representation capabilities.

## 5 Deep Learning for Pest Recognition

As discussed in the previous analysis of pest image datasets, images collected from different sources exhibit diverse characteristics, including variations in target scale, background complexity, species distribution, and environmental conditions. These differences introduce several challenges for intelligent pest identification, such as small-object detection, dense target recognition, complex background interference, and domain variation. Meanwhile, traditional machine learning approaches rely heavily on handcrafted features and therefore show limited adaptability to these complex scenarios. To address these challenges, deep learning approaches have gradually become the dominant paradigm in pest identification due to their powerful capability of automatically learning hierarchical visual representations from large-scale image data [39].

With the rapid development of deep learning, image recognition based on Convolutional Neural Networks (CNNs) has achieved major breakthroughs. Its adaptive feature extraction capability can accurately capture key characteristics of pests such as morphology and texture, laying a solid technical foundation for intelligent pest identification [40]. Through multi-layer non-linear transforms, CNNs learn deep semantic representations automatically and are able to adapt to the characteristics of small body sizes, varied morphologies, and strong adhesion to the background of crop leaves during the application of field crop-related visual tasks, demonstrating clear technical advantages.

Since around 2016, convolutional neural networks (CNNs) have been increasingly applied to pest identification, achieving substantially higher recognition accuracy than conventional image processing and machine learning techniques and showing stronger adaptability to the complexity and diversity of field environments [41,42]. On the basis of CNN architectures, a series of image classification networks have

been developed, including AlexNet [43], VGG [44], GoogLeNet [45], ResNet [46], and EfficientNet [47]. AlexNet, VGG, GoogLeNet, and ResNet achieved excellent performance in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), whereas EfficientNet, which was introduced after the competition ended, later demonstrated strong performance on the ImageNet classification task. These achievements demonstrate the powerful feature extraction capabilities of deep convolutional networks. Although these architectures were originally designed for image-level classification tasks, their learned feature representations have also provided important foundations for subsequent object detection frameworks through transfer learning and backbone design. However, these models were initially developed for general image classification scenarios, which differ substantially from agricultural pest identification tasks involving complex backgrounds, multi-scale targets, and large intra-class variations. Therefore, task-specific optimization is required before they can be effectively applied to pest identification.

Beyond pure image classification, the precise localization and identification of pests in complex field environments—such as determining the specific positions of pests on crop leaves and stems to provide a basis for precise pesticide application—mainly rely on object detection networks. Models derived from CNNs have shown robust recognition performance on well-known public benchmark datasets including Pascal VOC and COCO [48,49], yet these benchmarks differ substantially from real-world agricultural pest datasets and cannot be directly transferred. From a practical application perspective, Pascal VOC and COCO mainly target general objects such as vehicles, animals, and daily necessities, with relatively uniform object scales, simple and neat backgrounds, and a small number of objects per image. In contrast, agricultural pest datasets have unique field properties: there is a large diversity of pests and each type of crop has an associated type of pest, and even the same crop may experience simultaneous infestations by several pests; pests can vary tremendously in size, both in millimeter-sized eggs and nymphs or centimeter-sized adults; they are densely packed, with more than one pest being found on a given leaf; the backgrounds are intricate and highly intertwined with crop tissues, whereby aphids are camouflaged with leaf textures and the tunnels of leaf miners are hard to differentiate between leaf veins. At the same time, the datasets are exposed to the influence of environmental parameters, i.e., the changing field lighting, the growth cycle of crops, and the level of pest infestations that cause increased variability in the distribution of data.

At present, dedicated object detection models including R-CNN [50], Fast R-CNN [51], Faster R-CNN [52], YOLO [53], and SSD [54] have been extensively explored in agricultural pest detection, all with targeted improvements tailored to the above characteristics of pest datasets. New-generation visual detection algorithms can be mainly divided into two categories: two-stage object detection frameworks represented by the R-CNN family, and one-stage detection strategies typified by YOLO and SSD.

Two-stage models have been intuitively depicted in Fig. 8 as having a sequential logic of region proposal generation, feature extraction, and classification and bounding box refinement. They first generate candidate boxes and then perform category judgment and coordinate correction. Though they are more computationally intensive, they produce better recognition results which are especially useful when high-precision is necessary, in the initial pest surveillance and hidden pest attack identification, to buy time on the active control of pests and infections. In contrast, one-stage object detection models adopt an end-to-end training and prediction process. While their recognition accuracy is generally slightly lower than that of two-stage models, their inference speed is drastically improved, rendering them more suitable for real-time field scenarios requiring rapid detection, such as drone-based field inspection and large-scale rapid pest screening, to enhance the efficiency of regional pest monitoring.

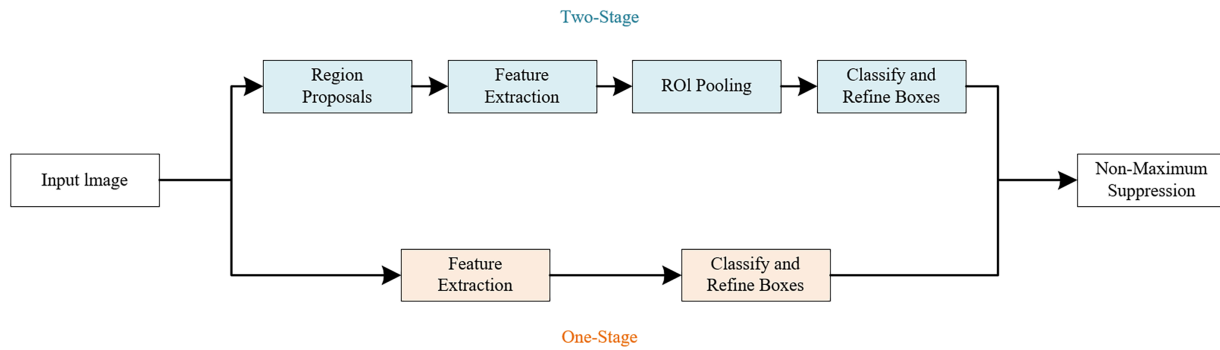


**Figure 8:** Flowchart of one-stage and two-stage recognition algorithms.

A systematic comparison of the core logic, key steps, representative algorithms, and performance characteristics of the two architectures is provided in Table 3. Selecting an appropriate detection model according to the actual scenarios of field pest monitoring is the key to achieving accurate and efficient pest surveillance.

**Table 3:** Comparison of one-stage and two-stage recognition architectures.

| Architecture Type | Core Mechanism                                                                                       | Representative Algorithms           | Advantages                                                                                               | Limitations                                                  |
|-------------------|------------------------------------------------------------------------------------------------------|-------------------------------------|----------------------------------------------------------------------------------------------------------|--------------------------------------------------------------|
| Two-Stage Model   | Generate region proposals first and then perform classification and bounding box refinement          | Faster R-CNN, Mask R-CNN            | Higher detection accuracy; strong localization capability; suitable for complex scenes and small targets | High computational cost; relatively slow inference speed     |
| One-Stage Model   | Directly predict object categories and bounding boxes without independent region proposal generation | YOLO Series (v5/v8), SSD, RetinaNet | Fast inference; end-to-end prediction; suitable for real-time deployment                                 | Accuracy may be lower for extremely small or complex targets |

### 5.1 One-Stage Networks

Considering the challenges of large scale variations, dense distributions, and real-time monitoring requirements in pest image datasets, one-stage detection frameworks have attracted extensive attention due to their balance between detection accuracy and inference efficiency. One-stage deep convolutional neural network (CNN)-based crop pest recognition methods jointly perform pest category classification and bounding box regression based on image features to achieve automatic pest identification, characterized by end-to-end training and real-time detection capabilities. Unlike two-stage frameworks that require separate region proposal generation, one-stage models directly predict pest categories and bounding box coordinates from input images, making them highly suitable for field real-time monitoring scenarios integrated with agricultural Internet of Things (IoT) devices [55].

The application of one-stage models in agriculture has followed a clear evolutionary trajectory, with early studies laying the foundation for subsequent advances. In 2016, Ding and Taylor [56] proposed an automatic deep learning-based pipeline for detecting and counting pests in farmland traps. A key contribution of this early work was to demonstrate the feasibility of real-time inference on parallel hardware without extensive manual engineering, providing an important proof of concept for field deployment. Nevertheless, shallow network architectures have limited capacity to learn the fine-grained texture information required to distinguish pests from complex vegetation backgrounds. Fuentes et al. [57] addressed these limitations in 2017 by developing a robust real-time detector for tomato pests and diseases using mainstream one-stage architectures and deeper feature extraction networks, including VGG and ResNet. This study established an important design principle: agricultural pest detectors require deeper feature hierarchies than generic object detectors because pest-induced symptoms can be visually similar to abiotic stress symptoms. The combination of local-global category labeling and active data augmentation further indicated that reducing false positives is as important as maintaining high recall in field applications. Subsequent studies have continued to address three interrelated challenges specific to agricultural scenarios: extreme scale variation across pest developmental stages, severe background clutter caused by leaves and soil, and strict computational constraints for battery-powered edge deployment.

The YOLO (You Only Look Once) series has emerged as the most widely adopted one-stage framework in pest detection, primarily because its architectural philosophy, simultaneous localization and classification, aligns with the operational tempo of field scouting. As illustrated in Fig. 9, its typical architecture consists of three core components, comprising a Backbone for feature extraction, a Neck for multi-scale feature aggregation, and a Head for task-specific prediction. This modular design has proven particularly amenable to agricultural customization, as each component can be independently optimized to address specific field challenges without destabilizing the overall detection pipeline.

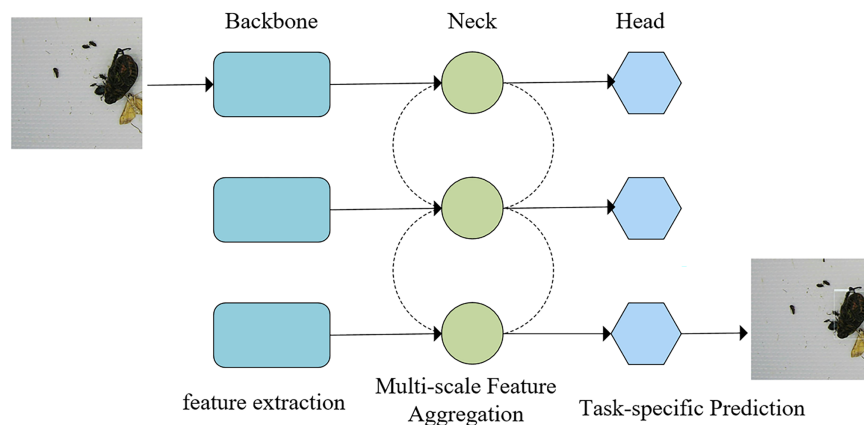


**Figure 9:** YOLO object detection model architecture.

For early agricultural adaptation and scale challenges, Partel et al. [58] employed YOLOv1 and YOLOv3 to monitor the Asian citrus psyllid (*Diaphorina citri*), eliminating cumbersome manual feature engineering through a dual-network detection and classification cascade. However, this application also revealed a fundamental mismatch between generic anchor box designs and the biological characteristics of pest populations. Adult psyllids measure only 3–4 mm and exhibit high intra-class morphological variation depending on wing posture and lighting conditions. Rigid statistical anchor boxes inherited from natural-image benchmarks are therefore poorly suited to such variability. To improve scale adaptation, Wang [59] proposed DS-YOLOv3, which incorporated a lightweight DenseNet backbone and a Spatial Pyramid Pooling module to enlarge the

effective receptive field. The Spatial Pyramid Pooling module was not merely an incremental improvement; rather, it was designed to reduce information loss caused by aggressive downsampling of small pests through consecutive convolutional layers. By preserving multi-scale contextual information, the model achieved a detection accuracy of 82.3% for minute and occluded pests, highlighting the importance of scale-aware architectural modifications in agricultural vision.

With the continuous development of YOLO architectures, research has diverged into two complementary optimization directions: enhancing representation capacity in complex backgrounds and achieving extreme lightweighting for edge deployment. This divergence reflects the consensus that a single model often struggles to meet both the high-accuracy requirements of early warning systems and the low-latency constraints of real-time spray controllers. To improve discriminative representation for morphologically similar pests in cluttered canopies, Yang et al. [60] proposed Maize-YOLO, which integrates CSP ResNeXT-50 and VoVG CSP modules and achieved 76.3% average precision on the challenging IP102 dataset. This result indicates that inter-stage partial connections, originally developed for general image classification, can be repurposed to improve gradient flow in networks that must distinguish subtle textural differences between pest cuticles and leaf epidermis. Because IP102 contains 102 fine-grained pest classes and severe class imbalance, it provides a more ecologically realistic benchmark than highly curated image collections; therefore, performance improvements on this dataset may be more relevant to operational field applications. To support deployment on resource-constrained IoT devices, Zha et al. [61] presented YOLOv4\_MF, a lightweight forest pest detector based on MobileNetV2 and coordinate attention. This strategy reflects the practical reality that continuous edge operation may be more important than marginal accuracy gains in many agricultural settings. Replacing standard convolution with depthwise separable convolution reduces both parameter size and inference energy consumption, factors that are often underrepresented in academic evaluations but substantially influence practical adoption. Studies published in 2024 further validated this dual pathway, as MACNet [62] and an optimized YOLO\_MRC [63] introduced content-aware feature sampling to dynamically balance model size and accuracy according to input complexity. Content-adaptive computation departs from static model design by allocating computational resources according to scene complexity, making it particularly suitable for agricultural environments where pest density and background complexity can vary considerably within a single field.

With the emergence of YOLO11, the requirements for lightweight deployment and accurate tiny-object detection in complex environments have become increasingly prominent. To address the deployment constraints of edge devices in smart granaries, Hou et al. [64] proposed ABF-YOLO11. By combining an ADown downsampling module with BiFPN, this model reduced the number of parameters to 1.50M while maintaining a high detection rate of 94.4% for stored-grain pests. This result indicates that aggressive parameter reduction does not necessarily compromise multi-scale feature fusion. The BiFPN design enables efficient cross-scale information transfer with low computational cost. From an agricultural engineering perspective, a model with only 1.5M parameters may enable device-side inference on low-cost microcontrollers, thereby improving the economic feasibility of large-scale pest monitoring deployments. To address very small pest size and cluttered backgrounds in American ginseng storage, Zhao et al. [65] proposed OS-YOLO. By introducing Omni-Dimensional Dynamic Convolution and a dedicated Tiny Object Detection Layer, OS-YOLO reduced background noise and improved fine-grained feature extraction, achieving an mAP of 96.2%. The Tiny Object Detection Layer explicitly addresses a key limitation of feature hierarchies optimized for medium-sized objects: such hierarchies often fail to preserve millimeter-scale target information. OS-YOLO mitigates this bias by assigning dedicated parameters to shallow, high-resolution feature maps. These findings suggest that effective agricultural detectors must simultaneously maintain low computational cost and accurate multi-scale detection, a dual requirement that is less common in conventional computer vision tasks.

Although the YOLO series is widely used for real-time agricultural detection, the Single Shot MultiBox Detector (SSD) framework represents another important branch of one-stage detection models. SSD is known for its strategy of extracting multi-scale features. As shown in Fig. 10, the standard SSD architecture extracts feature maps from different network depths and directly predicts bounding boxes using multi-scale detection heads, followed by standard class-wise Non-Maximum Suppression (NMS). This multi-scale prediction mechanism, based on feature maps of different resolutions, anticipated the need to process visual patterns at different object scales separately and has since influenced many detection frameworks. Deng et al. [66] applied this capability to stored-grain pest detection and adjusted the loss function to improve training efficiency. However, performance plateaued at 78.5% mAP because the model remained insufficiently adapted to extremely small targets. This limitation reflects a structural weakness of SSD: shallow feature maps used for small-object prediction receive weaker training signals and contain less semantic information, making it difficult to distinguish pest-like debris from actual infestations. The core bottleneck is therefore not only detection sensitivity but also the limited semantic representation of early-layer features. To address this issue, Guo et al. [67] proposed a scale-adaptive SSD variant for bee recognition, using image splitting and NMS to improve small-target localization. Although image splitting improves the effective resolution of small targets, it may also increase inference time and divide large pests across tile boundaries. This trade-off explains why no single architectural choice is universally optimal for all agricultural detection applications.

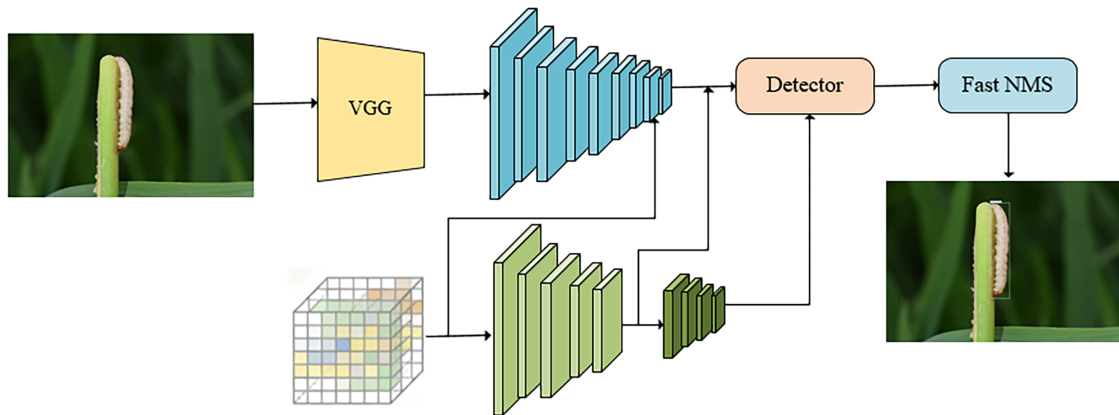


**Figure 10:** SSD model architecture.

With the development of research, one-stage frameworks have been moving towards anchor-free and end-to-end paradigms in order to overcome the shortcomings of strict anchor-based design in the initial versions of YOLO and SSD. This tendency is illustrated by Zhang et al. [68] who suggested the Pest-RTDetr, an optimized anchor-free architecture based on improved cross-scale feature fusion and improved loss functions. With the removal of complex post-processing processes like Non-Maximum Suppression, this simplified version was able to obtain a 68.4% mAP at 80 FPS on the IP102 dataset, which is very efficient in detecting dense and multi-scale pests. The removal of Non-Maximum Suppression is especially important in the agricultural context. When pests are densely clustered, like aphids, conventional NMS tends to miss detections because it suppresses any overlapping bounding boxes. NMS assumes that repeated detections of the same thing are extraneous, but on the farm there can be several neighboring individuals that are truly separate targets. Anchor-free and NMS-free structures bypass this assumption altogether and provide a more ecologically realistic detection model to gregarious pest species.

To provide a comprehensive overview of the aforementioned one-stage architectures, Table 4 summarizes the performance metrics, core technological optimizations, and target application scenarios of these representative models, enabling a direct comparative analysis of their evolutionary efficacy. It should be noted that the quantitative results reported in different studies cannot be directly compared due to variations in experimental settings. Specifically, the reviewed methods were evaluated on different pest datasets with different category numbers, image acquisition conditions, annotation standards, and task definitions. Moreover, differences in input image resolution, data augmentation strategies, training schedules, hardware platforms, and evaluation protocols may also affect the reported performance. Therefore, the performance values summarized in the tables are intended to illustrate the technical characteristics and development trends of different approaches rather than provide a strict ranking among models.

Despite significant progress, one-stage models still face challenges: static anchor box designs struggle with diverse pest morphologies, shallow feature fusion limits tiny pest detection, and existing datasets often lack coverage of emerging invasive species [69]. Future improvements should focus on dynamic anchor box generation, adaptive multi-scale feature fusion, and few-shot learning for rare pests. Additionally, lightweight optimization through depthwise separable convolution and model pruning will facilitate integration with low-power IoT devices for widespread field application [70].

**Table 4:** Performance of representative one-stage crop pest recognition methods.

| Model Name                           | Technologies                                             | Target Scenario/Pest Type                  | Accuracy/mAP      |
|--------------------------------------|----------------------------------------------------------|--------------------------------------------|-------------------|
| YOLOv1 [58]                          | Basic YOLO architecture                                  | Citrus orchard pest monitoring             | 80% (accuracy)    |
| SSD (VGG Backbone) [57]              | Multi-scale feature fusion and data augmentation         | Tomato pest detection                      | 85.98% (mAP)      |
| DS-YOLOv3 [59]                       | Lightweight backbone and multi-scale feature enhancement | Small and occluded pest detection          | 82.3% (accuracy)  |
| Maize-YOLO [60]                      | Feature extraction and fusion optimization               | Maize pest detection                       | 76.3% (mAP)       |
| YOLOv4_MF [61]                       | Lightweight network with attention mechanism             | Forest pest detection                      | 88.93% (mAP)      |
| MACNet [62]                          | Adaptive feature sampling and lightweight design         | General crop pest detection                | 95.73% (accuracy) |
| YOLO_MRC [63]                        | Lightweight optimization                                 | Tephritidae pest detection                 | 71.2% (mAP)       |
| SSD (Stored-Grain Pest Version) [66] | Multi-scale feature maps, loss function optimization     | Stored-grain pest detection                | 78.5% (mAP)       |
| Scale-Adaptive SSD [67]              | Scale adaptation and localization optimization           | Complex environment pest detection         | 89.9% (accuracy)  |
| CA-YOLOv5s [71]                      | Attention-based feature enhancement                      | Maize leaf pest detection                  | 48.6% (mAP)       |
| Swin Transformer-YOLOv5m [72]        | Transformer-based feature modeling                       | General crop pest detection                | 95.7% (accuracy)  |
| AC-YOLO [73]                         | Attention mechanism and loss optimization                | Fine-grained stored-grain pest recognition | 91.9% (mAP)       |

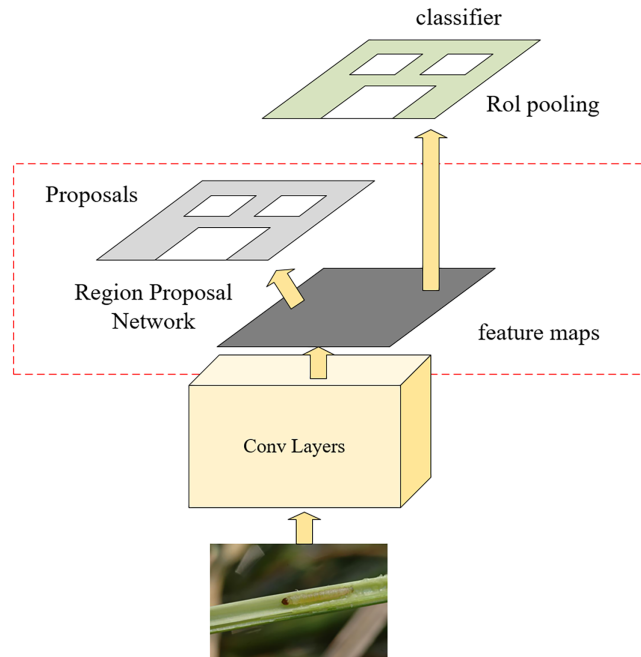
(Continued)

**Table 4 (continued)**

| Model Name            | Technologies                                  | Target Scenario/Pest Type                       | Accuracy/mAP |
|-----------------------|-----------------------------------------------|-------------------------------------------------|--------------|
| YOLOv8-FasterTea [74] | Transfer learning strategy                    | Tea pest detection under data-scarce conditions | 71.8% (mAP)  |
| Improved YOLOv8n [75] | Feature representation and multi-scale fusion | Small-target flax pest detection                | 94.5% (mAP)  |

## 5.2 Two-Stage Networks

Although one-stage detectors provide efficient end-to-end prediction, their performance may still be challenged by dense pest distributions, severe occlusion, and complex backgrounds commonly observed in practical agricultural environments. Therefore, two-stage frameworks have been explored to improve localization and classification accuracy by introducing region proposal mechanisms. Two-stage deep convolutional neural network-based crop pest recognition methods improve detection accuracy for small or partially occluded targets by narrowing the search space to regions of interest (ROIs) [16,76]. These models consist of two relatively independent stages—region proposal network (RPN) and detection network—providing high flexibility in module design while enabling in-depth feature learning and classification optimization [77]. Compared with one-stage detection methods, region proposal-driven frameworks improve detection accuracy by focusing computational resources on potential target areas, demonstrating superior performance in complex agricultural scenarios with cluttered backgrounds or overlapping pests [78,79]. Fig. 11 illustrates the core architecture of Faster R-CNN.

**Figure 11:** Faster R-CNN model architecture.

Early work on two-stage pest detection mostly focused on basic testing of region proposal methods and different ways to get features. But these first tries already showed that there is a natural conflict between how detailed computation can get and what actually works well when used in farming. In 2016, Liu et al. [76]

proposed a two-stage deep learning method for pest localization and classification in rice field images. Their method first generated saliency maps by calculating the global contrast of image regions to identify potential pest targets, and then cropped and resized the detected regions to construct standardized pest image samples for subsequent model training. On this basis, they employed a deep convolutional neural network (DCNN) to allow the model to autonomously learn local image features from this dataset, thereby achieving hierarchical abstraction from simple low-level features to complex high-level patterns, and finally accomplish correct classification. This study presents a truly feasible two-stage detection process, while also pointing out an existing problem: although the region proposals based on significance are computationally inexpensive, they are prone to errors when used in actual field conditions. Fields contain plenty of debris that looks like pests, as well as old leaf spots that also resemble the targets. This generates a large number of false positive results and severely interferes with detection. The heavy reliance on low-level contrast cues created a bottleneck, prompting researchers to develop deeper, learning-based methods for region proposal generation. Back in 2017, Yang et al. [79] managed to make computations faster by using image color attenuation. They combined spatial interactions between superpixel areas to work out saliency values for places where pests might be, and also made target localization and segmentation more precise with the GrabCut algorithm. An improved version of AlexNet was used to get features from the segmented pest images and sort them into different classes. This method made it easier to adapt to complex backgrounds, but it also showed a clear basic problem with pipelines that rely on segmentation. The GrabCut refinement step does help make boundary measurements more accurate, but it adds extra iterative optimization steps. These steps are hard to fit with the real-time speed needs that monitoring systems used out in the field have to meet. How to achieve both improved classification results and shorter inference time—this balance issue is often overlooked by those who attempt to deploy two-stage agricultural detectors.

The Faster R-CNN framework is now a key base for two-stage pest detection. This is mostly because it splits region proposal and classification into separate modules, which matches the two-step thinking process that agricultural experts use. They first look over areas where pests might be living, then do a close, specific ID of what kind of pest it is. In 2018, Shen et al. [80] made a dataset covering six common species of stored-grain insects. This dataset included images of insects mixed in with grains and other impurities, and all insect locations were marked by hand with bounding boxes. They also built a Faster R-CNN model that added an improved Inception network to work as the feature extractor. In this setup, the RPN made candidate regions that might hold insects, and the classification network did the work of identifying the species and adjusting the bounding box positions. The team chose an Inception-based backbone on purpose, to deal with the fact that stored-grain pests come in many different sizes. Pest sizes can be as small as tiny beetles or as large as bigger moths, and this big size range causes problems for shallower networks that have fixed receptive fields. That study showed that two-stage frameworks separate localization work from classification work, so people can use specialized feature extractors that are made to fit the size distribution of the target pest group. Li et al. [81] put forward a two-stage method for automatic finding and counting crop pests back in 2019. In this method, RPN together with Non-Maximum Suppression (NMS) was used to generate ROIs, and a changed ZFCNN worked for pest detection. Experimental results demonstrated that this method achieved an average simulation accuracy of 0.885. However, this excellent performance was achieved under controlled conditions. Although the ZFCNN architecture is computationally lightweight and responds quickly, its feature extraction depth is limited. When dealing with pests that are densely distributed and mutually occluded, there may be difficulties in accurately separating them. Therefore, when choosing a two-stage model, it is necessary to pay special attention to the gap between the simulation accuracy and the field practical robustness. To deal with the problems that come with small size, high density, and complicated backgrounds when studying brown planthoppers (*Nilaparvata lugens*), He et al. [82] made a new two-stage detector in 2020. In the first

step of this process, the RPN quickly finds the possible areas where planthoppers live near the base of rice plants, and then an improved detection network does exact identification in the second step. The test results show that this detector works better than single-stage models, and also better than the original YOLOv3. Planthoppers gather at the base of rice plants, where the background clutter is most confusing and the difference between the target and its background is the smallest. In these kinds of environments, the way two-stage models split resources from coarse to fine actually works better than the uniform computation that one-stage detectors use.

A defining limitation of traditional anchor-based two-stage detectors is the rigidity of predefined anchor box geometries, which are ill-equipped to accommodate the extreme ontogenetic and postural variation observed in field pest populations. To address this, recent optimizations have pivoted towards adaptive and anchor-free mechanisms. For instance, Dong et al. [16] proposed CRA-Net, which consists of a channel recalibration feature pyramid network (CRFPN) and an adaptive anchor (AA) module. The CRFPN captures discriminative features and improves the recognition accuracy and localization of small pests, while the AA module alleviates inefficient matching between anchors and ground-truth boxes. Together, these two components improve the effectiveness of multi-class small-pest detection. Pushing this further, Jiao et al. [77] proposed an Anchor-Free Region Proposal Network, completely eliminating manual anchor tuning and achieving highly reliable target proposals for 24 pest species on a large-scale dataset. The elimination of anchor hyperparameters addresses a practical barrier to agricultural adoption, namely the expertise and computational resources required to recalibrate anchor configurations for each new crop-pest system. Anchor-free region proposal thus enhances both generalization across diverse agricultural settings and accessibility for practitioners without deep learning specialization.

Beyond scale adaptation and anchor design, researchers have introduced various structural modifications to two-stage models to enhance fine-grained classification capability, which remains a major challenge in agriculture because many pest groups exhibit highly similar appearances. To distinguish closely related pest species, Chen et al. [83] developed a contextual feature enhancement module combined with adaptive multi-scale fusion to improve feature representation for morphologically similar pest categories. This design reduced the impact of suboptimal feature selection and achieved a mean average precision of 72.7% in their evaluation experiments. The contextual feature enhancement module addresses a specific agricultural problem: morphologically similar pests may occupy different microhabitats on host plants. By expanding the receptive field to include surrounding leaf context, the model can learn subtle habitat-related cues associated with species identity, introducing an ecological perspective that is absent from purely appearance-based classifiers. To address the irregular spatial distribution of pest damage on crops, Chen et al. [84] proposed TSAF-Net, an anchor-free rotated region proposal network that generates high-quality rotated bounding boxes without extensive hyperparameter tuning. Rotated bounding boxes are particularly useful in agricultural scenarios because pest damage patterns, such as leaf-miner tunnels and stem-borer entry holes, often align with leaf veins or stem axes. In contrast, axis-aligned boxes may include excessive background information, thereby weakening discriminative signals for subsequent classification. TSAF-Net demonstrates that biologically informed geometric modeling can improve detection accuracy. Another related approach, IIDC-GOARDL [85], introduces the Gannet optimization algorithm to automatically tune hyperparameters, thereby improving classification stability and reducing the manual effort required to adapt the model to new pest categories.

The inherent modularity of two-stage frameworks can also be extended naturally to pixel-level instance segmentation. This capability is increasingly important in precision agriculture because practical applications require not only detection of pest presence but also assessment of infestation severity and spatial distribution. Building on Mask R-CNN, He et al. [86] introduced a pixel reweighting strategy to accurately segment

silkworm bodies and feed edges in complex industrial environments. The transition from bounding-box detection to instance segmentation is nontrivial for agricultural data because pest boundaries are often ambiguous due to occlusion by host tissues or adhesion to debris. Pixel reweighting encourages the model to preserve the spatial continuity of silkworm pixels and distinguish their texture from feed material, incorporating a domain-specific bias that is not explicitly represented in general segmentation models. Comparative studies, such as that of Li et al. [87], suggest that single-stage models perform well in simple-background scenarios, whereas two-stage designs remain advantageous in complex agricultural environments with diverse species. Nevertheless, this two-stage paradigm incurs substantial computational cost and latency. This contradiction has prompted research on lightweight region proposal networks and model distillation. To provide a comparative overview of these improvements, Table 5 summarizes the key technologies, target scenarios, and performance metrics of representative two-stage models. The same consideration applies to the performance results summarized in Table 5, where reported metrics are mainly used to indicate methodological characteristics and research trends rather than direct numerical comparisons.

**Table 5:** Performance of representative two-stage crop pest recognition methods.

| Model Name                           | Technologies                                                    | Target Scenario/Pest Type                              | Accuracy/mAP              |
|--------------------------------------|-----------------------------------------------------------------|--------------------------------------------------------|---------------------------|
| Faster R-CNN [80]                    | Improved backbone and region proposal optimization              | Stored-grain pest detection                            | 88% (mAP)                 |
| ZF CNN [81]                          | ROI-based detection and CNN feature optimization                | Crop pest localization and counting                    | 85.98% (mAP)              |
| AF-RCNN [77]                         | Anchor-free proposal generation and feature fusion              | Multi-class crop pest detection                        | 56.4% (mAP)               |
| CRA-Net [16]                         | Feature pyramid enhancement and adaptive anchor optimization    | Small-scale pest detection                             | 67.9% (AP)                |
| Improved Mask R-CNN [86]             | Pixel-level refinement and boundary optimization                | Silkworm identification and counting                   | 96.83% (accuracy)         |
| Optimized Two-Stage R-CNN [78]       | Adaptive feature pyramid fusion and fine-grained detection      | Multi-class agricultural pest detection                | 77% (average accuracy)    |
| Two-Stage Similar-Pest Detector [83] | Context-aware feature enhancement and multi-scale fusion        | Fine-grained similar pest identification               | 72.7% (mAP)               |
| IIDC-GOALD [85]                      | Deep feature extraction and optimization-based parameter tuning | General agricultural pest detection and classification | 98.52% (average accuracy) |

Two-stage convolutional neural network-based detection methods possess significant advantages in handling pest detection tasks under complex backgrounds owing to their cascaded architecture of region proposal followed by fine classification. By integrating adaptive optimization strategies for anchor boxes, multi-scale feature pyramids, and attention mechanisms, researchers have achieved high recognition accuracy in detecting small and densely distributed pest targets under complex field scenarios. Nevertheless,

such high-precision performance often comes at the cost of reduced inference speed and increased computational resource consumption, and when confronted with extreme scale variations among pests and highly morphologically similar species, traditional two-stage models still exhibit insufficient feature extraction fidelity and limited fine-grained recognition accuracy. Future research should therefore concentrate on two core challenges, multi-scale detection and fine-grained identification in complex field environments. For the extreme size range spanning from aphids to armyworms, it is necessary to further optimize feature extraction and scale adaptation capabilities, minimize the loss of small-target features caused by aggressive downsampling, and reduce the rate of missed detection for tiny pests. Concurrently, for closely related pests of the same family or camouflaged pests that are highly mimetic to crop backgrounds, the network's capacity to attend to key discriminative anatomical regions should be strengthened to enhance the discriminability of fine-grained features. Addressing these challenges will break through the limitations of current models in recognizing similar targets and further enhance the adaptability and practical utility of two-stage models in real field scenarios.

### 5.3 Vision Transformers

The application of Transformer models to crop pest detection has brought about a major shift in the research thinking within the field of agricultural vision. These models are not based on the traditional convolution-based frameworks but incorporate self-attention mechanisms and encoder-decoder architectures as shown in Fig. 12. Self-attention mechanism allows efficient modeling of long-range dependencies between pests and their intricate environments. At the same time, the end-to-end parallel prediction design can maintain competitive detection accuracy while simplifying the detection pipeline by removing the region proposal mechanisms commonly used in two-stage models. All these architectural benefits make Transformer models particularly well suited to tackle innate agricultural problems, including extremely cluttered backgrounds, intense pest occlusion, and very strong multi-scale variation of targets.

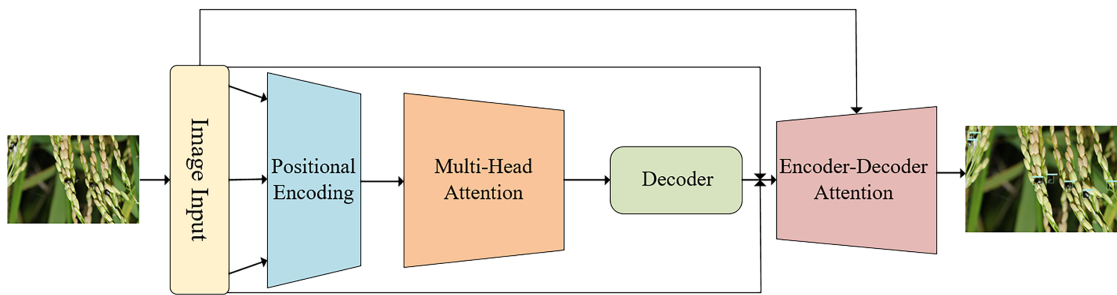


**Figure 12:** Transformer model architecture.

Computer vision development of Transformers provided the basis behind applications within agriculture. First introduced as a part of natural language processing [88], the self-attention mechanism has been extended to vision using Non-local Neural Networks [89] and then DETR [90], the first end-to-end detector based on Transformer. Following improvements, like the Swin Transformer [91] and Vision Transformer (ViT) [92] proved to be more versatile in visual tasks than ever before, making it possible to use them in more sophisticated pest detection systems. It should be noted that these general visual Transformers are trained and validated on common natural image datasets, where objects have similar scales and simple backgrounds. Such models cannot be directly applied to agricultural pest detection. Given the unique characteristics of agricultural scenarios, targeted adjustments and optimizations to the network structure are required to meet practical detection requirements.

The initial studies of the application of Transformer to agriculture included two mutually reinforcing approaches, which were the combination of convolutional local feature extraction and Transformer global attention, and the use of cross-modal information to alleviate the lack of annotated agricultural data. Peng and Wang [93] integrated the CNN backbone model to extract local features with the Transformer attention head to refine global features with high precision without dynamic feature alignment in serial fusion. Although the serial CNN-Transformer architecture is easy to implement, it has a fundamental weakness, i.e., the convolutional front-end acts independently of the Transformer global context modeling. Morphologically similar pests that inhabit different ecological niches might benefit out of joint local-global reasoning, a functionality that serial architectures can only partially support. Such an observation encourages the development of more tightly integrated hybrid systems. Lu et al. [94] designed a Ghost-convolution enlightened Transformer to overcome resource limitations using lightweight Ghost convolutions to achieve an impressive rate of 98.14 percent accuracy in grape pest dataset. The lightweight design meets a critical deployment issue, namely, that conventional Transformer blocks have a quadratic computational complexity depending on the input resolution which becomes prohibitively expensive when processing high-resolution agricultural images where pests typically take up only a small number of pixels. The burden is alleviated by ghost convolutions as they create redundant feature maps using cheap linear transformations hence maintaining representational capacity at reduced computational costs. Simultaneously, multi-modal and self-supervised methods started to be used in order to decrease the data dependence. ITF-WPI was introduced by Dai et al. [95] to creatively combine image and text features to recognize the wolfberry pests, whereas MMFGT was introduced by Zhang et al. [96] to combine self-supervised learning with fine-grained analysis. These multi-modal initiatives are especially relevant to agricultural settings where textural descriptions of pest morphology, host preference and seasonality offer rich supplementary cues lacking in pixel data alone. The combination of visual and textual modalities therefore allows a kind of knowledge-enhanced perception not possible with pure vision models.

With the growth of the Transformer study, the focus has now been placed on addressing particular performance bottlenecks and increasing model deployability in field environments. Tang et al. [97] improved Pest-YOLO by adding ECA mechanism and Transformer encoder, reaching a 73.4% mAP on the Pest24 dataset. Combining the efficient channel attention mechanism with Transformer encoders holds great research value. This combination can effectively make up for the deficiencies of traditional self-attention and solve the problem of fine-grained feature loss during global information aggregation. Maintaining channel-wise feature fidelity is important in pest detection because the task relies on detecting very small textural differences in the cuticle or wing venation. The ECA module enhances informative channels prior to the application of global attention. Venkatasachandran and Iyapparaja [98], in order to validate the effectiveness of transfer learning in custom-made architectures, created a GNViT model based on ViT to classify groundnuts pests specifically. The emergence of ViT variants tailored for specific crops is due to the fact that the general ImageNet-style pre-training often performs poorly when transferred to the agricultural field, and the visual statistical characteristics of different crops vary greatly. Finetuning ViT on crop-specific pest imagery enables the self-attention mechanism to reweight patch embeddings relative to agriculturally meaningful patterns, like leaf venation intersections where pests tend to gather. To ensure even stronger model robustness, Xie and Ye [99] proposed SFA-ViT with multi-factor data augmentation and multi-scale attention. The strong emphasis on robustness via aggressive data augmentation is entirely reasonable in agricultural applications where imaging environment is uncontrolled and diverse. Conventional augmentation pipelines trained on natural images are frequently unable to replicate the very particular degradations experienced in the field, such as motion blur due to wind-induced canopy movement

or specular highlights on waxy leaves. Multi-factor augmentation that directly captures these agricultural imaging artifacts produces more resilient models.

Recent advances have pushed the boundaries of feature aggregation and low-data adaptation, two frontiers of particular relevance to agricultural pest detection. A 2025 study proposed Pest-ConFormer, a hybrid CNN-Transformer model featuring a dual-path structure and weakly-supervised fine-grained modules, which achieved a 77.81% accuracy on the IP102 dataset, outperforming state-of-the-art methods by 2 percentage points [100]. The dual-path architecture is conceptually significant because it formalizes a division of labor between local texture analysis, for which convolutions are ideally suited, and global contextual reasoning, for which self-attention excels. In agricultural imagery, a pest's identity is often jointly determined by fine-scale cuticular patterns and coarse-scale host plant association, a duality that single-path architectures struggle to reconcile. Addressing severe low-data scenarios, a 2025 study developed the MMFSL model, which fuses ViT-extracted visual features with pre-trained language model text embeddings. Under an extreme 5-way 1-shot setting, it reached an impressive 86.97% accuracy on PlantVillage and 56.78% on complex field datasets [101]. The stark performance differential between PlantVillage and complex field datasets, 86.97% vs. 56.78%, is diagnostically revealing. It underscores that even sophisticated few-shot learners remain vulnerable to the domain gap between curated laboratory images and unconstrained field photography. Bridging this gap will require not only algorithmic advances but also the curation of few-shot benchmarks that faithfully reflect field deployment conditions.

While Transformer-based models have achieved remarkable success, their practical application in agricultural settings is constrained by several inherent limitations that are often underappreciated in the computer vision literature. The substantial computational overhead restricts seamless deployment on resource-constrained edge devices. This constraint is particularly acute in agriculture, where edge devices may be solar-powered and situated in remote locations with intermittent connectivity. The latency and energy costs of Transformer inference can exceed the operational budgets of such devices by orders of magnitude. Furthermore, global self-attention mechanisms frequently fail to extract the fine-grained features of minute pests. This limitation stems from the quadratic scaling of attention with respect to the number of input tokens. When processing high-resolution imagery to preserve small-pest details, the resulting token sequences become prohibitively long, forcing practitioners to adopt patch sizes that effectively downsample away the very signals they seek to detect. Moreover, their robust performance remains highly dependent on the availability of large-scale annotated datasets. In agricultural domains, large-scale annotation is seldom feasible due to the expertise required for fine-grained pest identification and the seasonal, ephemeral nature of pest outbreaks. The resulting data scarcity creates a fundamental tension with the data-hungry nature of Transformer architectures. Moving forward, overcoming these barriers requires a strategic shift toward lightweight optimizations specifically tailored to agricultural edge hardware, such as sparse attention mechanisms that restrict computation to agriculturally salient regions and neural architecture search that co-optimizes for accuracy and energy efficiency. Additionally, integrating self-supervised pre-training on abundant unlabeled agricultural data will be essential for reducing dependence on manual annotation, as the volume of unlabeled field imagery collected by IoT devices and drones far exceeds what can feasibly be labeled. Advancing cross-modal fusion paradigms that incorporate structured agricultural knowledge, such as pest-host phenology and regional distribution maps, will further enhance model robustness and interpretability. These directions will be pivotal in translating Transformer-based models from theoretical promise into scalable, real-world agricultural solutions. To provide a concise comparison of these Transformer-based studies, Table 6 summarizes their representative architectures, key technological strategies, target application scenarios, and reported performance metrics.

**Table 6:** Performance of representative transformer-based crop pest recognition methods.

| Model Name                        | Technologies                                                             | Target Scenario/Pest Type                                  | Accuracy/mAP                                         |
|-----------------------------------|--------------------------------------------------------------------------|------------------------------------------------------------|------------------------------------------------------|
| CNN-Transformer Hybrid Model [93] | CNN-Transformer hybrid feature modeling                                  | General crop pest recognition                              | 99.47% and 97.94% on the D0 dataset and Li's dataset |
| GeT [94]                          | Lightweight Transformer architecture                                     | Grape pest detection under resource-constrained conditions | 98.14% (accuracy)                                    |
| ITF-WPI [95]                      | Vision-language feature fusion                                           | Wolfberry pest recognition                                 | 97.98% (accuracy)                                    |
| MMFGT [96]                        | Self-supervised Transformer representation learning                      | General crop pest recognition under low-data conditions    | 98.12% (accuracy)                                    |
| Improved Pest-YOLO [97]           | Attention-enhanced Transformer detection                                 | Real-time crop pest detection                              | 73.4% (mAP)                                          |
| SFA-ViT [99]                      | Multi-scale Transformer attention and data augmentation                  | Crop pest and disease recognition                          | 74.70% (accuracy)                                    |
| Pest-ConFormer [100]              | CNN-Transformer hybrid architecture and fine-grained feature aggregation | Multi-class crop pest recognition                          | 77.81% (accuracy)                                    |
| MMFSL [101]                       | Multimodal few-shot vision-language learning                             | Few-shot crop pest and disease recognition                 | 86.97% (accuracy)                                    |

## 6 Multimodal Large Models for Pest Identification and Diagnosis

Traditional deep learning methods have achieved significant success in the image classification and object detection of agricultural pests. However, they typically output only single category labels or bounding boxes, lacking the capacity for logical reasoning regarding pathological causes and the ability to generate specific prevention and control measures. In recent years, the rapid emergence of Large Language Models (LLMs) and Vision-Language Models (VLMs) has brought a paradigm shift to the field of agricultural pest management. These large models can not only process complex cross-modal data but also leverage massive prior knowledge bases and powerful logical reasoning capabilities to achieve a complete closed loop—from ‘simple recognition’ to ‘intelligent diagnosis and prescription generation’.

### 6.1 Decoupled Pipeline Architecture of Visual Perception and Large Language Models

In decoupled systems, pest detection and identification are still performed primarily by specialized computer vision models, whereas the LLM operates on structured recognition outputs rather than raw image pixels. Typical inputs to the LLM include the predicted pest category, confidence score, host crop, detected location, damage symptoms, and environmental observations. The language model then integrates these recognition results with domain knowledge to verify the predicted category, distinguish visually similar pests, explain the diagnostic basis, and provide subsequent management information. Despite their substantial

potential, directly deploying large models is often impractical in agricultural scenarios constrained by limited edge-device computing power. Therefore, many studies adopt a decoupled architecture that combines a lightweight front-end visual model with a cloud-based back-end large model. A schematic diagram of this decoupled diagnostic process is shown in Fig. 13.

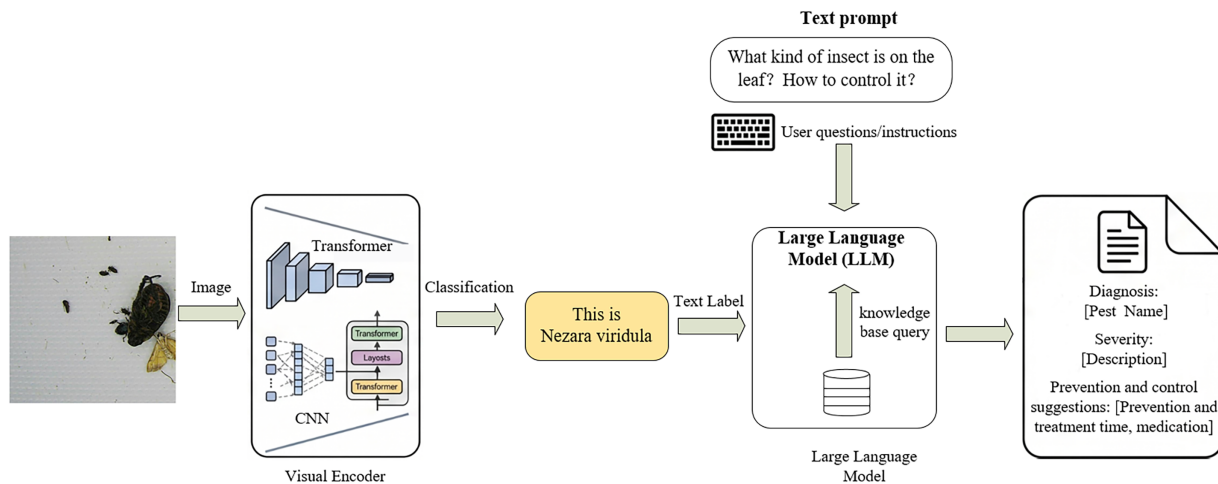


**Figure 13:** Dual-model decoupled diagnostic architecture.

Yan et al. [102] proposed the CDIP-ChatGLM3 framework, which employs an efficient computer vision model (EfficientNet-B2) at the front end to extract visual features and identify crop diseases, and then converts the recognition results into diagnostic keywords. Subsequently, these keywords are input into a localized ChatGLM3-6B large language model, which has undergone dual-stage mixed fine-tuning (DMT) using professional agricultural books, to further integrate domain knowledge and generate structured diagnostic information. By separating visual recognition from language-based reasoning, this approach reduces the computational burden of end-to-end multimodal models while improving the interpretability of recognition results, making it suitable for practical agricultural applications. Similarly, Qing et al. [103] designed an edge-cloud collaborative diagnostic system based on lightweight YOLOPC and GPT-4. This method utilizes an improved YOLOPC model at the edge side to perform visual feature extraction and pest/disease recognition, and converts the detected visual characteristics into natural language descriptions through an Image-to-Text Mapping mechanism. Following this, GPT-4 combines these visual descriptions with external information, such as orchard geographical location and climate conditions, to perform semantic reasoning and generate comprehensive diagnostic reports. In this framework, the visual model remains responsible for target recognition, while GPT-4 provides knowledge integration and interpretation based on recognition results.

## 6.2 End-to-End Agricultural Multimodal Large Language Models

With the maturation of multimodal feature alignment technologies, end-to-end agriculture-specific MLLMs, which directly integrate image and text information within a unified framework, have emerged as a promising research direction in intelligent pest identification (as illustrated in Fig. 14). By establishing effective connections between visual representations and linguistic knowledge, these models can improve the semantic understanding of pest images and support fine-grained recognition, open-vocabulary identification, and the analysis of visually similar pest categories. Compared with traditional decoupled frameworks, end-to-end MLLMs provide stronger cross-modal reasoning capabilities, enabling more flexible

interpretation of pest-related visual information while maintaining the advantages of language-based knowledge integration.

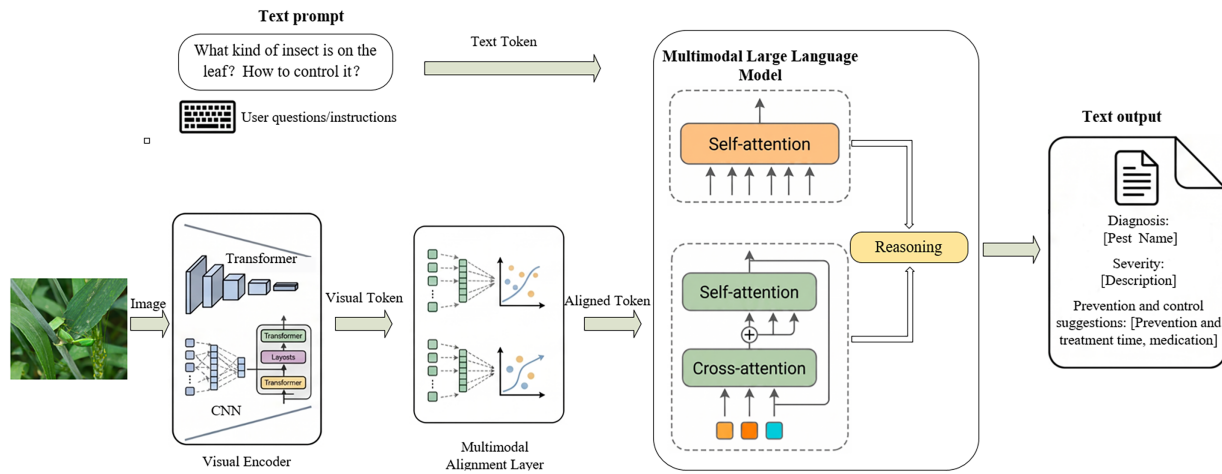


**Figure 14:** End-to-end multimodal VLM-based intelligent diagnosis architecture.

Wang et al. [104] proposed a multimodal large model, LLMI-CDP, based on the VisualGLM architecture. To address the issues of “hallucinations” and the lack of domain knowledge typical of general large models in agriculture-specific scenarios, this study introduced the Q-Former framework as a bridge in the model architecture. This framework efficiently maps the image features extracted by the vision encoder into the semantic space of the large language model, achieving precise alignment between visual features and linguistic semantics. Furthermore, considering the computational deployment thresholds for implementing agricultural large models, the model employed Low-Rank Adaptation (LoRA) technology, achieving efficient fine-tuning while freezing the vast majority of pre-trained parameters. This end-to-end design enables the model to move beyond outputting single disease categories; instead, it outputs scientific prevention and control recommendations—including detailed symptom analysis and specific medication guidance—in natural language through multi-turn dialogue interactions.

Although LLMI-CDP has demonstrated promising performance, challenges remain in visual feature extraction and cross-modal semantic alignment when dealing with complex farmland backgrounds, visually similar pest categories, and small-scale pest targets. To address these limitations, Ahmedm and Luan proposed [105] AgriM-LLM, a multimodal large model specifically optimized for agricultural visual understanding. The main contributions of this research are reflected in the progressive optimization of three core modules. First, to address the large variations in pest morphology and complex visual appearances, the model designed an Enhanced Vision Encoder (EVE) incorporating a Multi-Scale Feature Fusion (MSFF) module. This mechanism effectively integrates local details with global contextual information, enabling the model to capture fine-grained characteristics of small pest targets as well as broader spatial patterns in agricultural images, thereby providing richer visual representations for distinguishing visually similar pests. Second, during the cross-modal alignment stage, the model introduced an Agriculture-Knowledge-Enhanced Q-Former (AKE-Q), which incorporates structured agricultural knowledge into query representations. This knowledge-guided attention mechanism helps the model focus on pest-related visual features while reducing interference from complex field backgrounds, improving the alignment between visual information and agricultural semantic concepts. Finally, the backend Domain-Adaptive Language Model (DALM) adopted a multi-stage progressive fine-tuning strategy. Through additional training on professional agricultural texts, the model acquired domain-specific terminology and reasoning

capabilities, enabling more accurate interpretation of visual recognition results and generation of reliable diagnostic explanations. These language-based outputs can further support agricultural decision-making as downstream applications of pest identification.

### 6.3 Cross-Modal Retrieval and Incremental Learning Based on Vision-Language Models

Compared with conventional closed-set recognition models, cross-modal retrieval and continual learning provide new solutions for pest identification under long-tailed distributions, limited annotations, and emerging pest categories. By aligning pest images with textual descriptions, these methods establish shared semantic representations between visual features and agricultural knowledge, enabling more flexible recognition, retrieval, and incremental updating of pest categories.

In natural farmland environments, pest images are often accompanied by highly complex background interference, and different species often exhibit the phenomenon of “high inter-class similarity and large intra-class variance”. Traditional pure visual recognition and retrieval systems struggle to accurately distinguish these highly similar pest categories due to a lack of semantic depth. To address this issue, Yu et al. [106] proposed the CPCPR cross-modal retrieval framework. The innovation of this study lies in utilizing the powerful natural language understanding and generation capabilities of large language models to automatically generate fine-grained text descriptions—including visual characteristics, morphological descriptions, and developmental stages—for the complex IP102 image dataset, thereby constructing a high-quality multimodal image-text paired dataset named the Cross-modal Crop Pest Retrieval Dataset (CCPRD). Regarding the model architecture, CPCPR employs a dual-tower vision-language model and innovatively designs a dual Progressive Attention Encoder (PAE). Through Vision Instruction Representation (VIR), this encoder can automatically calculate a confidence matrix, intelligently filter out irrelevant background noise, and force the model to focus on the salient pest-related regions. Furthermore, to tackle the problem where highly similar species are prone to confusion in the feature space, the framework introduces a clustering-based Affiliation Loss. This effectively enhances the compactness of intra-class pest samples and improves the discrimination of visually similar pest categories in the cross-modal semantic space, achieving robust image-text alignment and precise pest retrieval under complex farmland backgrounds.

In practical agricultural production, the ecological environment is highly dynamic, and new pest species continuously emerge with changes in seasons and climate. When learning the features of new pests, traditional deep learning models often forget the previously mastered knowledge of old pests—a phenomenon known as “catastrophic forgetting”—which severely restricts the long-term deployment of recognition systems. To endow the model with continuous learning capabilities, Hu et al. [107] proposed the PestCLIP incremental recognition framework based on the CLIP large model. This method abandons traditional pure visual incremental fine-tuning in favor of a dual-prompt tuning strategy. By combining fixed text templates with learnable context vectors, it adapts to new pest tasks while maximizing the retention of the VLM’s original generalized knowledge through regularization. More ingeniously, to review old knowledge without significantly increasing the storage burden on edge devices, the framework constructs a feature-level “Concept Pool”. Instead of storing massive amounts of historical raw images, it extracts old pest categories as statistical features of Gaussian distributions for storage and sampling replay, achieving highly efficient knowledge consolidation. In addition, to counter the long-tail bias that pre-trained large models are prone to exhibit on niche agricultural datasets, PestCLIP introduces a Prediction Distribution Calibration (PDC) mechanism based on causal analysis. This mechanism performs post-alignment of the output logits, thereby maintaining excellent incremental recognition accuracy even in the face of continuously emerging unknown pests.

Beyond cross-modal retrieval and incremental learning, recent vision-language research has begun to extend multimodal perception toward knowledge-enhanced reasoning and decision support in precision agriculture. To address the hallucination problems that may arise during the practical deployment of large models, researchers have increasingly integrated vision-language models with Internet of Things (IoT) sensors and external agricultural knowledge bases. The PEZEGO intelligent precision agriculture system, developed by Yuan et al. [108], employs the HCLORA method to fine-tune an open-vocabulary detector for precise zero-shot pest detection. At the decision-making stage, the system introduces a Structured Data-Based Retrieval-Augmented Generation (SRAG) workflow that combines professional knowledge represented by the large model with real-time environmental data, such as temperature and humidity, collected by IoT devices. Through multi-step logical reasoning based on Chain of Thought (CoT), PEZEGO further interprets visual recognition results and associates them with agricultural knowledge to generate reliable diagnostic information and management recommendations. Although its primary objective is agricultural decision support rather than cross-modal retrieval or incremental learning, this system demonstrates how vision-language representations can be connected with external knowledge retrieval and structured environmental information, thereby extending multimodal pest identification from visual perception to interpretable decision-making in precision agriculture.

## 7 Discussions

The development of intelligent crop pest identification technologies reflects a continuous transition from handcrafted feature engineering toward automated visual perception and multimodal intelligence. Traditional machine learning approaches established the foundation of automated pest recognition by reducing dependence on manual identification. By extracting handcrafted features such as color, texture, and shape information, these methods achieved satisfactory performance under relatively controlled environments and were suitable for early-stage pest recognition tasks with limited training samples. However, compared with deep learning-based approaches, traditional machine learning methods have limited representation capability because their performance strongly depends on manually designed features and predefined statistical assumptions. Consequently, their adaptability remains insufficient in complex agricultural scenarios involving illumination variations, background interference, pest posture changes, and morphological similarities among different species.

Compared with traditional machine learning approaches, deep learning methods have significantly improved pest identification by enabling automatic hierarchical feature learning from large-scale image datasets. CNN-based models, particularly object detection frameworks, have enhanced the capability to recognize small-scale pests, dense pest populations, and visually similar categories. These advantages make deep learning the dominant technical paradigm in current pest identification research. However, the improvement in recognition performance is accompanied by increased requirements for annotated data, computational resources, and training complexity. Moreover, most existing deep learning-based studies still focus on optimizing model performance on specific benchmark datasets, while their generalization ability across different geographical regions, crop growth stages, environmental conditions, and image acquisition devices remains insufficient. Therefore, high accuracy achieved under laboratory or benchmark conditions does not necessarily indicate reliable performance in real agricultural environments.

Another critical challenge originates from the limitations of existing pest datasets. Although current datasets have greatly promoted the development of intelligent identification algorithms, differences in data acquisition strategies, including Internet-based collection, pest monitoring systems, and field image acquisition, inevitably introduce sampling bias and domain variation. Internet-collected datasets usually provide clear pest appearances and rich category diversity, which are beneficial for species recognition but may not fully represent complex field conditions. Monitoring-lamp datasets support continuous surveillance

and dense pest analysis, but their distributions are influenced by specific environmental conditions and pest behavioral characteristics. Field-collected datasets provide more realistic agricultural scenarios but are often constrained by high acquisition costs, annotation difficulties, and insufficient sample diversity. Therefore, existing benchmark datasets may not fully reflect the complexity of practical pest management scenarios. Future intelligent pest identification systems should pay more attention to multi-source dataset construction, cross-domain adaptation, and evaluation under realistic field conditions.

Vision Transformers and multimodal large models provide new opportunities for advancing intelligent pest identification from visual recognition toward agricultural knowledge-based understanding. Compared with CNN-based models that mainly focus on local feature extraction, Vision Transformer-based approaches improve the modeling of global visual relationships through attention mechanisms, providing enhanced capability for capturing complex pest appearance variations. Furthermore, compared with conventional recognition models that only output pest categories or localization results, multimodal large models integrate visual information with language-based agricultural knowledge and have the potential to support more comprehensive functions, including pest diagnosis, damage assessment, and management recommendation. Nevertheless, these emerging approaches are still in an early stage of agricultural application. Challenges such as hallucination problems, insufficient domain-specific knowledge, high computational requirements, limited interpretability, and the lack of standardized evaluation protocols remain major obstacles for practical deployment.

In addition to algorithmic performance, deployment capability is another important factor determining the practical value of intelligent pest identification systems. Most existing studies evaluate models using offline datasets and high-performance computing platforms, whereas real agricultural applications require consideration of inference speed, memory consumption, energy efficiency, and hardware compatibility. Although lightweight networks and edge computing technologies provide potential solutions, maintaining a balance between accuracy and efficiency remains challenging, especially for complex tasks involving small targets, dense distributions, and dynamic environmental changes. Moreover, agricultural environments are highly variable due to differences in weather conditions, crop growth stages, geographic regions, and pest populations. Therefore, future intelligent pest identification systems should not only pursue higher recognition accuracy but also achieve efficient deployment, continuous adaptation, and reliable operation under diverse field conditions.

Overall, intelligent crop pest identification is gradually evolving from isolated visual recognition models toward integrated intelligent agricultural systems combining perception, reasoning, and decision support. Despite significant progress achieved by traditional machine learning, deep learning, Vision Transformer, and multimodal large model approaches, challenges related to data diversity, model generalization, knowledge integration, and practical deployment remain unresolved. Addressing these issues will be essential for developing reliable and intelligent pest management technologies and provides the foundation for the future research directions discussed in the following section.

## 8 Future Perspectives

Although intelligent pest recognition technologies have achieved substantial progress, their widespread deployment in real-world agricultural production remains limited. Recent studies have explored improvements from multiple perspectives, including data resource construction, model optimization, multimodal information fusion, adaptive learning, and intelligent decision support. These studies have demonstrated the potential of artificial intelligence technologies to improve pest monitoring efficiency and support precision plant protection. However, significant challenges remain in terms of data diversity, model generalization, computational efficiency, and practical applicability. Therefore, future research should continue to advance

from multiple dimensions, including data resource construction, model optimization, multimodal integration, system deployment, and ecological-level application, in order to establish more accurate, reliable, and intelligent pest monitoring systems.

- (1) Recent studies have constructed various pest datasets through different acquisition strategies, including Internet-based collection, pest monitoring systems, and field image acquisition. These datasets have greatly promoted the development of intelligent pest recognition algorithms and provided important benchmarks for model evaluation. However, current pest datasets still face challenges related to limited sample size, insufficient category coverage, long-tailed distribution, and domain differences between training and practical deployment environments. Future work should overcome the inherent limitations of existing pest data collection by developing data systems that closely integrate key agronomic characteristics, including crop growth stages, pest geographical distribution, seasonal occurrence patterns, and damage severity. Unified standards for image acquisition, annotation, and quality control should be formulated. With the support of efficient data augmentation approaches, such as automatic annotation and generative adversarial networks (GANs), samples of rare pests, newly emerging invasive pests, and pests at different instars and damage levels should be systematically supplemented. This will facilitate the construction of large-scale open-source datasets covering wider geographical regions, complete seasonal cycles, and all developmental stages. Such datasets would help address long-tailed distributions and domain shifts, reduce the negative impact of data imbalance on model generalization, and provide a stronger data foundation for field deployment of intelligent identification technologies.
- (2) Existing studies have investigated different object detection architectures, including two-stage and one-stage detection frameworks, as well as lightweight model optimization strategies such as depthwise separable convolution, knowledge distillation, and model pruning. These approaches have achieved promising results in reducing computational complexity while maintaining acceptable recognition performance. However, balancing detection speed and recognition accuracy remains a core challenge for practical deployment. Under complex field conditions, such as uneven illumination, dense canopy occlusion, and small pest targets, current models still struggle to achieve both high accuracy and low inference latency, particularly on low-computing-power and energy-constrained hardware such as unmanned aerial vehicles (UAVs) and agricultural Internet of Things (IoT) nodes. Future model optimization should focus on lightweight and adaptive technologies. Combined with dynamic anchor box generation, adaptive feature fusion, and dynamic scale adaptation, these strategies can support the development of edge-computing models with compact model sizes and strong environmental adaptability. While maintaining real-time monitoring capability, such models can improve detection accuracy for tiny and overlapping pests, enable seamless integration on low-power IoT devices and UAVs, and support high-frequency field inspection.
- (3) Recently, transfer learning, domain adaptation, few-shot learning, and foundation model-based approaches have been explored to improve the adaptability of pest recognition models under limited data conditions. These methods provide potential solutions for recognizing rare pest categories and reducing dependence on large-scale annotated datasets. However, existing approaches are still mainly evaluated under relatively controlled experimental conditions, and their robustness under complex agricultural environments remains insufficient. Farmland ecosystems are characterized by continuous dynamic evolution driven by crop growth stage succession, field biological community succession, regional climate variations, and cropping system changes. Such dynamics frequently lead to the occurrence of sudden and regional emerging pests as well as alien invasive pests, which impose stringent requirements on adaptive recognition capability. Meanwhile, domain shifts caused

by differences in ecological zones, crop types, cultivation patterns, and field environments further restrict the generalization and application of models in practical crop protection monitoring. Therefore, future research should focus on constructing robust cross-domain transfer learning and incremental learning frameworks in line with agronomic rules. Relying on the generalized representation capability and universal knowledge of large-scale pre-trained models such as CLIP, combined with few-shot and zero-shot learning mechanisms, models can rapidly adapt to monitoring scenarios across different production regions, crop types, and growth stages with only a small number of annotated samples or even without annotated samples. These strategies can improve the capability of identifying emerging and sudden pest species while avoiding catastrophic forgetting of existing pest knowledge, thereby meeting the requirements of real-time monitoring and emergency prevention and control in field plant protection.

- (4) Recent advances in multimodal learning and large vision-language models have stimulated increasing interest in agricultural applications. Existing studies have explored the integration of visual recognition models with language models to provide pest diagnosis, knowledge retrieval, and agricultural decision support. These studies indicate that multimodal approaches have the potential to extend pest identification beyond simple visual recognition toward intelligent agricultural systems. However, current multimodal pest recognition methods remain in an early development stage, and challenges such as insufficient agricultural domain knowledge, unreliable reasoning results, and limited integration with field environmental information still need to be addressed. Future intelligent recognition of crop pests will break through the limitations of single visual recognition and evolve toward multi-sensory collaborative perception integrated with multi-source habitat information. Based on actual production demands in field plant protection, it is essential to establish intelligent pest identification and decision-making systems with multimodal information fusion. Such systems should deeply integrate pest images with core agronomic and ecological environmental data, including meteorological factors such as temperature, humidity, precipitation, and sunlight, crop canopy hyperspectral characteristics, soil physicochemical properties, and soil moisture conditions. By taking advantage of complementary multi-source information, these systems can improve the robustness and accuracy of pest recognition under complex field environments and establish a full-chain framework from pest identification and population dynamics prediction to scientific prevention and control strategy generation.
- (5) Explainable artificial intelligence (XAI) and knowledge-enhanced reasoning methods have recently attracted increasing attention in agricultural applications. Existing studies have attempted to improve model transparency through attention visualization, knowledge graphs, and reasoning mechanisms. However, most current pest recognition models still mainly focus on prediction accuracy and lack sufficient explanation of their decisions, which limits their acceptance in practical agricultural management. Future research should enhance the explainability and decision transparency of pest recognition models. For field plant protection and agricultural technology extension, intelligent pest recognition models must overcome the black-box limitations of traditional algorithms and strengthen agronomic-level interpretability. By introducing plant protection knowledge graphs and reasoning mechanisms based on pest occurrence patterns, models can not only identify pest species using image features but also explain agronomic evidence such as pest morphological characteristics, host feeding habits, and field outbreak drivers. When further combined with professional plant protection knowledge bases, crop growth progress, field ecological conditions, and pest population dynamics, models can generate evidence-based green prevention and control recommendations suitable for practical production. This transformation from visual feature extraction to semantic reasoning integrated with agronomic mechanisms can improve the professionalism and credibility of intelligent recognition results and promote the development of intelligent plant protection toward a more scientific, accurate, and field-oriented direction.

## 9 Conclusions

This paper reviews the evolution of intelligent crop pest identification technologies from traditional machine learning and deep learning to multimodal large models. At the data level, the review analyzes the inherent sampling biases of three data sources—pest monitoring lamps, manual field collection, and web crawling—revealing resulting problems such as long-tailed distributions, domain shifts across scenes, and the absence of critical agronomic attributes. The findings highlight that single-source data alone can hardly support stable model generalization across regions and seasons. At the methodology level, the paper successively reviews the applicable boundaries and performance bottlenecks of traditional methods including clustering, support vector machines, and decision trees. Subsequent sections systematically compare the differential performance of one-stage, two-stage, and Transformer-based deep learning paradigms in terms of accuracy, speed, and field robustness. Furthermore, the analysis explores how multimodal large models, through cross-modal semantic alignment and agricultural knowledge reasoning, extend pest identification from mere visual classification to a closed-loop service chain encompassing “identification–diagnosis–decision-making,” marking a paradigm shift from instrumental detection to knowledge-driven intelligent plant protection.

Nevertheless, several core challenges in this field remain to be addressed. First, although existing datasets continue to grow in scale, they generally lack key agronomic annotations such as crop growth stage, geographical eco-region, pest population density, and damage severity, hindering the reasoning required for plant protection decisions like “when to control” and “whether the economic threshold has been reached.” Second, the generalization ability of most models across different ecological zones and cropping systems remains weak; when confronted with newly invasive species or regionally emerging pests, these models often require large amounts of freshly labeled data for retraining and still lack the capability for rapid adaptation under low-shot or even zero-shot conditions. Moreover, a gap persists between current identification systems and subsequent control actions—model outputs are mostly limited to category labels or bounding boxes, failing to achieve closed-loop integration with precision spraying decisions, optimal timing of control, and efficacy assessment. Addressing these issues will be the key to advancing intelligent pest identification from technical verification to large-scale field deployment.

**Acknowledgement:** The authors would like to thank all contributors and institutions that supported this research.

**Funding Statement:** The authors received no specific funding for this study.

**Author Contributions:** The authors confirm contribution to the paper as follows: Conceptualization, Shuai Zhou and Zhigui Liu; methodology, Shuai Zhou and Jun Wu; literature search and analysis, Shuai Zhou and Li Li; investigation, Shuai Zhou, Rong Tang, Zhangjun Peng and Mingfei Wan; formal analysis, Shuai Zhou and Zhiqiang Chen; resources, Zhigui Liu; literature organization, Shuai Zhou and Li Li; writing—original draft preparation, Shuai Zhou; writing—review and editing, Zhigui Liu; visualization, Shuai Zhou; supervision, Zhigui Liu; project administration, Zhigui Liu. All authors reviewed and approved the final version of the manuscript.

**Availability of Data and Materials:** Not applicable.

**Ethics Approval:** This study did not involve human or animal participants and therefore did not require ethical approval.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. National agro-technical extension and service center: 23 major crop pests and diseases to occur in a severe outbreak trend [Internet]. [cited 2026 Aug 31]. Available from: <http://www.icama.org.cn/zwb/detail/29756>.
2. FAO. The impact of disasters and crises on agriculture and food security. Rome, Italy: Food and Agriculture Organization of the United Nations; 2021.
3. Xiao Z, Yin K, Geng L, Wu J, Zhang F, Liu Y. Pest identification via hyperspectral image and deep learning. *Signal Image Video Process*. 2022;16(4):873–80. doi:10.1007/s11760-021-02029-7.
4. Xie J, Lu M, Gao Q, Chen L, Zou Y, Wu J, et al. Intelligent detection and control of crop pests and diseases: current status and future prospects. *Agronomy*. 2025;15(6):1416. doi:10.3390/agronomy15061416.
5. Guo Y, Gao J, Wang X, Jia H, Wang Y, Zeng Y, et al. Precious tree pest identification with improved instance segmentation model in real complex natural environments. *Forests*. 2022;13(12):2048. doi:10.3390/f13122048.
6. Wang D, Deng L, Ni J, Gao J, Zhu H, Han Z. Recognition pest by image-based transfer learning. *J Sci Food Agric*. 2019;99(10):4524–31. doi:10.1002/jsfa.9689.
7. Guan B, Wu Y, Zhu J, Kong J, Dong W. GC-faster RCNN: the object detection algorithm for agricultural pests based on improved hybrid attention mechanism. *Plants*. 2025;14(7):1106. doi:10.3390/plants14071106.
8. Chen CJ, Huang YY, Li YS, Chang CY, Huang YM. An AIoT based smart agricultural system for pests detection. *IEEE Access*. 2020;8:180750–61. doi:10.1109/ACCESS.2020.3024891.
9. Gao D, Sun Q, Hu B, Zhang S. A framework for agricultural pest and disease monitoring based on Internet-of-Things and unmanned aerial vehicles. *Sensors*. 2020;20(5):1487. doi:10.3390/s20051487.
10. Nyakuri JP, Nkundineza C, Gatera O, Nkurikiyeyezu K, Mwitende G. AI and IoT-powered edge device optimized for crop pest and disease detection. *Sci Rep*. 2025;15(1):22905. doi:10.1038/s41598-025-06452-5.
11. Liu L, Wang R, Xie C, Yang P, Wang F, Sudirman S, et al. PestNet: an end-to-end deep learning approach for large-scale multi-class pest detection and classification. *IEEE Access*. 2019;7:45301–12. doi:10.1109/ACCESS.2019.2909522.
12. Liu J, Wang X. Plant diseases and pests detection based on deep learning: a review. *Plant Methods*. 2021;17(1):22. doi:10.1186/s13007-021-00722-9.
13. Wu X, Zhan C, Lai YK, Cheng MM, Yang J. IP102: a large-scale benchmark dataset for insect pest recognition. In: *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2019 Jun 15–20; Long Beach, CA, USA. p. 8779–88. doi:10.1109/cvpr.2019.00899.
14. Liu Y, Liu S, Xu J, Kong X, Xie L, Chen K, et al. Forest pest identification based on a new dataset and convolutional neural network model with enhancement strategy. *Comput Electron Agric*. 2022;192:106625. doi:10.1016/j.compag.2021.106625.
15. Xie C, Wang R, Zhang J, Chen P, Dong W, Li R, et al. Multi-level learning features for automatic classification of field crop pests. *Comput Electron Agric*. 2018;152:233–41. doi:10.1016/j.compag.2018.07.014.
16. Dong S, Wang R, Liu K, Jiao L, Li R, Du J, et al. CRA-Net: a channel recalibration feature pyramid network for detecting small pests. *Comput Electron Agric*. 2021;191:106518. doi:10.1016/j.compag.2021.106518.
17. Huang ML, Chuang TC. A database of eight common tomato pest images. *V1 Mendeley Data*. 2020. doi:10.17632/s62zm6djd2.1.
18. Wang QJ, Zhang SY, Dong SF, Zhang GC, Yang J, Li R, et al. Pest24: a large-scale very small object data set of agricultural pests for multi-target detection. *Comput Electron Agric*. 2020;175:105585. doi:10.1016/j.compag.2020.105585.
19. Li W, Wang D, Li M, Gao Y, Wu J, Yang X. Field detection of tiny pests from sticky trap images using deep learning in agricultural greenhouse. *Comput Electron Agric*. 2021;183:106048. doi:10.1016/j.compag.2021.106048.
20. Rustia DJA, Chao JJ, Chiu LY, Wu YF, Chung JY, Hsu JC, et al. Automatic greenhouse insect pest detection and recognition based on a cascaded deep learning classification method. *J Appl Entomol*. 2021;145(3):206–22. doi:10.1111/jen.12834.
21. Wang K, Li W, Wu X, Xu J, Wang ZW, Yang S. AP162: a large-scale dataset for agricultural pest recognition. *Comput Electron Agric*. 2025;237:110520. doi:10.1016/j.compag.2025.110520.

22. Watts MJ, Worner SP. Estimating the risk of insect species invasion: Kohonen self-organising maps versus k-means clustering. *Ecol Model.* 2009;220(6):821–9. doi:10.1016/j.ecolmodel.2008.12.016.
23. Li Z, Hong T, Zeng X, Zheng J. Citrus red mite image target identification based on K-means clustering. *Trans Chin Soc Agric Eng.* 2012;28(23):147–53. doi:10.3969/j.issn.1002-6819.2012.23.020.
24. Wang Z, Wang K, Zhang S, Liu Z, Mu C. Whiteflies counting with K-means clustering and ellipse fitting. *Trans Chin Soc Agric Eng.* 2014;30(1):105–12. doi:10.3969/j.issn.1002-6819.2014.01.014.
25. Liu G, Zhang Y, Huang J, Tang W. A method for image segmentation and recognition of spider mites based on K-means clustering algorithm. *Acta Entomol Sin.* 2015;58(12):1338–43. doi:10.16380/j.kcxb.2015.12.010.
26. García J, Pope C, Altimiras F. A distributed K-means segmentation algorithm applied to *Lobesia botrana* recognition. *Complexity.* 2017;2017:5137317. doi:10.1155/2017/5137317.
27. Han R. Research on rapid detection and recognition of farmland pests based on machine vision [dissertation]. Hangzhou, China: Zhejiang University; 2014. (In Chinese).
28. Ebrahimi MA, Khoshtaghaza MH, Minaei S, Jamshidi B. Vision-based pest detection based on SVM classification method. *Comput Electron Agric.* 2017;137:52–8. doi:10.1016/j.compag.2017.03.016.
29. Yang G. Research on recognition and detection of early nymphs of *Oxya chinensis* based on machine vision [dissertation]. Hangzhou, China: Zhejiang University; 2017. (In Chinese).
30. Barman U, Pathak C, Mazumder NK. Comparative assessment of Pest damage identification of coconut plant using damage texture and color analysis. *Multimed Tools Appl.* 2023;82:1–23. doi:10.1007/s11042-023-14369-2.
31. Yao Q, Chen GT, Wang Z, Zhang C, Yang BJ, Tang J. Automated detection and identification of white-backed planthoppers in paddy fields using image processing. *J Integr Agric.* 2017;16(7):1547–57. doi:10.1016/S2095-3119(16)61497-1.
32. Chen X, He W, Ye Z, Gai J, Lu W, Xing G. Soybean seed pest damage detection method based on spatial frequency domain imaging combined with RL-SVM. *Plant Methods.* 2024;20(1):130. doi:10.1186/s13007-024-01257-5.
33. Shen R, Zhang H, Chai B, Wang W, Wang G, Yan B, et al. BAFL-SVM: a blockchain-assisted federated learning-driven SVM framework for smart agriculture. *High Confid Comput.* 2025;5(1):100243. doi:10.1016/j.hcc.2024.100243.
34. Ye YW. Machine learning based monitoring and counting algorithm for major vegetable pests and its application system [dissertation]. Guangzhou, China: South China Agricultural University; 2018. (In Chinese).
35. Bu JY, Sun GX, Wang YX, Wei TX, Wang X. Identification and monitoring method of pests in greenhouse tomato crops based on sticky trap images. *J Nanjing Agric Univ.* 2021;44(2):373–83. (In Chinese).
36. Yang Z, Li W, Li M, Yang X. Automatic greenhouse pest recognition based on multiple color space features. *Int J Agric Biol Eng.* 2021;14(2):188–95. doi:10.25165/j.ijabe.20211402.5098.
37. Bütüner AK, Şahin YS, Erdinç A, Erdoğan H, Lewis E. Enhancing pest detection: assessing *Tuta absoluta* (Lepidoptera: Gelechiidae) damage intensity in field images through advanced machine learning. *J Agr Sci-Tarim Bili.* 2024;30(1):99–107. doi:10.15832/ankutbd.1308406.
38. Qiu L, Wu X, Xiong Q. Identification of pests and diseases of *Dalbergia hainanensis* based on EVI time series and classification of decision tree. *IOP Conf Ser Earth Environ Sci.* 2017;69(1):012162N.
39. Gao Z, Li Z, Zhang C, Wang Y, Su J. Double self-attention based fully connected feature pyramid network for field crop pest detection. *Comput Mater Contin.* 2025;83(3):4353–71. doi:10.32604/cmc.2025.061743.
40. Li Y, Wang H, Dang LM, Sadeghi-Niaraki A, Moon H. Crop pest recognition in natural scenes using convolutional neural networks. *Comput Electron Agric.* 2020;169:105174. doi:10.1016/j.compag.2019.105174.
41. Wu X. Research on pest recognition method based on machine vision [dissertation]. Hangzhou, China: Zhejiang University; 2016. (In Chinese).
42. Cheng X, Zhang Y, Chen Y, Wu YZ, Le Y. Application of deep learning in feature extraction and classification of stored grain pests. *J West Anhui Univ.* 2017;33(05):67–72. (In Chinese). doi:10.55446/ije.2024.1973.
43. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Commun ACM.* 2017;60(6):84–90. doi:10.1145/3065386.
44. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556.* 2014. doi:10.48550/arXiv.1409.1556.

45. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015 Jun 7–12; Boston, MA, USA. p. 1–9. doi:10.1109/cvpr.2015.7298594.
46. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. p. 770–8. doi:10.1109/CVPR.2016.90.
47. Tan M, Le QV. EfficientNet: rethinking model scaling for convolutional neural networks. *Int Conf Mach Learn*. 2019;97:6105–14.
48. Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A. The pascal visual object classes (VOC) challenge. *Int J Comput Vis*. 2010;88(2):303–38. doi:10.1007/s11263-009-0275-4.
49. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: common objects in context. In: Proceedings of the Computer Vision—ECCV 2014. Cham, Switzerland: Springer; 2014. p. 740–55. doi:10.1007/978-3-319-10602-1\_48.
50. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition; 2014 Jun 23–28; Columbus, OH, USA. p. 580–7. doi:10.1109/CVPR.2014.81.
51. Girshick R. Fast R-CNN. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). New York, NY, USA: The Association for Computing Machinery (ACM); 2015. p. 1440–8. doi:10.1109/ICCV.2015.169.
52. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans Pattern Anal Mach Intell*. 2017;39(6):1137–49. doi:10.1109/tpami.2016.2577031.
53. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. p. 779–88. doi:10.1109/CVPR.2016.91.
54. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, et al. SSD: single shot MultiBox detector. In: Proceedings of the Computer Vision—ECCV 2016. Cham, Switzerland: Springer; 2016. p. 21–37. doi:10.1007/978-3-319-46448-0\_2.
55. Cheng W, Liu Z, Li X. Enhanced lightweight architecture for real-time detection of agricultural pests and diseases. *Comput Mater Contin*. 2026;87(2):1–10. doi:10.32604/cmc.2025.074250.
56. Ding W, Taylor G. Automatic moth detection from trap images for pest management. *Comput Electron Agric*. 2016;123:17–28. doi:10.1016/j.compag.2016.02.003.
57. Fuentes A, Yoon S, Kim SC, Park DS. A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition. *Sensors*. 2017;17(9):2022. doi:10.3390/s17092022.
58. Partel V, Nunes L, Stansly P, Ampatzidis Y. Automated vision-based system for monitoring Asian *Citrus* psyllid in orchards utilizing artificial intelligence. *Comput Electron Agric*. 2019;162:328–36. doi:10.1016/j.compag.2019.04.022.
59. Wang W. Research on online recognition and targeted pesticide application of rice pests based on improved YOLOv3 [dissertation]. Hefei, China: Anhui Agricultural University; 2020. (In Chinese). doi:10.26919/d.cnki.gannu.2020.000012.
60. Yang S, Xing Z, Wang H, Dong X, Gao X, Liu Z, et al. Maize-YOLO: a new high-precision and real-time method for maize pest detection. *Insects*. 2023;14(3):278. doi:10.3390/insects14030278.
61. Zha M, Qian W, Yi W, Hua J. A lightweight YOLOv4-based forestry pest detection method using coordinate attention and feature fusion. *Entropy*. 2021;23(12):1587. doi:10.3390/e23121587.
62. Hu Y, Wang Q, Wang C, Qian Y, Xue Y, Wang H. MACNet: a more accurate and convenient pest detection network. *Electronics*. 2024;13(6):1068. doi:10.3390/electronics13061068.
63. Wei M, Zhan W. YOLO\_MRC: a fast and lightweight model for real-time detection and individual counting of Tephritidae pests. *Ecol Inform*. 2024;79:102445. doi:10.1016/j.ecoinf.2023.102445.
64. Hou L, Liu M, Liu Y, Tan X, Cheng C, Sun W, et al. ABF-YOLO11: a multi-scale adaptive detection and lightweight recognition model for stored-grain pests. *J Stored Prod Res*. 2026;116:102932. doi:10.1016/j.jspr.2025.102932.

65. Zhao C, Li S, Suthisut D, Bai C, Yan L, Wang D, et al. OS-YOLO: a robust object detection model for storage pest monitoring in American ginseng. *J Agric Food Res.* 2026;100521. doi:10.1016/j.jafr.2026.100521.
66. Deng Z, Wang P, Song X, Wang C, Chen J, Wu L. Research on grain granary pest detection based on SSD. *Comput Eng Appl.* 2020;56(11):214–8. (In Chinese). doi:10.2139/ssrn.4007163.
67. Guo X, Zhu Y, Li S, Zhang J, Lv C, Liu S. Scale adaptive small objects detection method in complex agricultural environment: taking bees as research object. *Smart Agric.* 2022;4(1):140–9. doi:10.12133/j.smartag.SA202203003.
68. Zhang L, Sun S, Zhao H, Li Z, Li D. Multi-species insect recognition method based on computer visions: sustainable agricultural development. *Ecol Inform.* 2025;88:103125. doi:10.1016/j.ecoinf.2025.103125.
69. Dong S, Du J, Jiao L, Wang F, Liu K, Teng Y, et al. Automatic crop pest detection oriented multiscale feature fusion approach. *Insects.* 2022;13(6):554. doi:10.3390/insects13060554.
70. Suzaudola M, Zhang D, Zeb A, Chen J, Wei L, Rayhan ABMS. Advanced deep learning model for crop-specific and cross-crop pest identification. *Expert Syst Appl.* 2025;274:126896. doi:10.1016/j.eswa.2025.126896.
71. Li Y, Sun S, Zhang C, Yang G, Ye Q. One-stage disease detection method for maize leaf based on multi-scale feature fusion. *Appl Sci.* 2022;12(16):7960. doi:10.3390/app12167960.
72. Dai M, Dorjoy MMH, Miao H, Zhang S. A new pest detection method based on improved YOLOv5m. *Insects.* 2023;14(1):54. doi:10.3390/insects14010054.
73. Zhao C, Bai C, Yan L, Xiong H, Suthisut D, Pobsuk P, et al. AC-YOLO: multi-category and high-precision detection model for stored grain pests based on integrated multiple attention mechanisms. *Expert Syst Appl.* 2024;255:124659. doi:10.1016/j.eswa.2024.124659.
74. Li K, Li Y, Wen X, Shi J, Yang L, Xiao Y, et al. Sticky trap-embedded machine vision for tea pest monitoring: a cross-domain transfer learning framework addressing few-shot small target detection. *Agronomy.* 2025;15(3):693. doi:10.3390/agronomy15030693.
75. Zhong M, Li Y, Gao Y. Research on small-target detection of flax pests and diseases in natural environment by integrating similarity-aware activation module and bidirectional feature pyramid network module features. *Agronomy.* 2025;15(1):187. doi:10.3390/agronomy15010187.
76. Liu Z, Gao J, Yang G, Zhang H, He Y. Localization and classification of paddy field pests using a saliency map and deep convolutional neural network. *Sci Rep.* 2016;6:20410. doi:10.1038/srep20410.
77. Jiao L, Dong S, Zhang S, Xie C, Wang H. AF-RCNN: an anchor-free convolutional neural network for multi-categories agricultural pest detection. *Comput Electron Agric.* 2020;174:105522. doi:10.1016/j.compag.2020.105522.
78. Jiao L, Xie C, Chen P, Du J, Li R, Zhang J. Adaptive feature fusion pyramid network for multi-classes agricultural pest detection. *Comput Electron Agric.* 2022;195:106827. doi:10.1016/j.compag.2022.106827.
79. Yang G, Bao Y, Liu Z. Localization and recognition of pests in tea plantation based on image saliency analysis and convolutional neural network. *Trans Chin Soc Agric Eng.* 2017;33(6):156–62. (In Chinese). doi:10.11975/j.issn.1002-6819.2017.06.020.
80. Shen Y, Zhou H, Li J, Jian F, Jayas DS. Detection of stored-grain insects using deep learning. *Comput Electron Agric.* 2018;145:319–25. doi:10.1016/j.compag.2017.11.039.
81. Li W, Chen P, Wang B, Xie C. Automatic localization and count of agricultural crop pests based on an improved deep learning pipeline. *Sci Rep.* 2019;9(1):7024. doi:10.1038/s41598-019-43171-0.
82. He Y, Zhou Z, Tian L, Liu Y, Luo X. Brown rice planthopper (*Nilaparvata lugens* Stal) detection based on deep learning. *Precis Agric.* 2020;21(6):1385–402. doi:10.1007/s11119-020-09726-2.
83. Chen H, Wang R, Du J, Chen T, Liu H, Zhang J, et al. Feature refinement method based on the two-stage detection framework for similar pest detection in the field. *Insects.* 2023;14(10):819. doi:10.3390/insects14100819.
84. Chen T, Chen H, Du J, Wang R, Zhang M, Zhang W. TSAF-Net: a rotated two-stage *Cnaphalocrocis medinalis* damage detection method based on anchor-free arbitrary-oriented proposal network. *Pest Manage Sci.* 2024;80(9):4604–16. doi:10.1002/ps.8184.
85. Al-Shahari EA, Aldehim G, Sharaf Almalki N, Assiri M, Sayed A, Alnfai MM. Innovative insect detection and classification for the agricultural sector using gannet optimization algorithm with deep learning. *IEEE Access.* 2024;12:108041–51. doi:10.1109/ACCESS.2024.3438308.

86. He R, Zheng K, Wei Q, Zhang X, Zhang J, Zhu Y, et al. Identification and counting of silkworms in factory farm using improved mask R-CNN model. *Smart Agric.* 2022;4(2):163–73. (In Chinese).
87. Li W, Zhu T, Li X, Dong J, Liu J. Recommending advanced deep learning models for efficient insect pest detection. *Agriculture.* 2022;12(7):1–17. doi:10.3390/agriculture12071065.
88. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. *Adv Neural Inf Process Syst.* 2017;30:5998–6008. doi:10.65215/r5bs2d54.
89. Wang X, Girshick R, Gupta A, He K. Non-local neural networks. In: *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*; 2018 Jun 18–23; Salt Lake City, UT, USA. p. 7794–803. doi:10.1109/CVPR.2018.00813.
90. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. In: *European Conference on Computer Vision (ECCV)*. Cham, Switzerland: Springer; 2020. p. 213–29.
91. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*; 2021 Oct 10–17; Montreal, QC, Canada. p. 9992–10002. doi:10.1109/iccv48922.2021.00986.
92. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16 × 16 words: transformers for image recognition at scale. *arXiv:2010.11929*. 2020. doi:10.48550/arXiv.2010.11929.
93. Peng Y, Wang Y. CNN and transformer framework for insect pest classification. *Ecol Inform.* 2022;72:101846. doi:10.1016/j.ecoinf.2022.101846.
94. Lu X, Yang R, Zhou J, Jiao J, Liu F, Liu Y, et al. A hybrid model of ghost-convolution enlightened transformer for effective diagnosis of grape leaf disease and pest. *J King Saud Univ Comput Inf Sci.* 2022;34(5):1755–67. doi:10.1016/j.jksuci.2022.03.006.
95. Dai G, Fan J, Dewi C. ITF-WPI: image and text based cross-modal feature fusion model for wolfberry pest recognition. *Comput Electron Agric.* 2023;212:108129. doi:10.1016/j.compag.2023.108129.
96. Zhang Y, Chen L, Yuan Y. Multimodal fine-grained transformer model for pest recognition. *Electronics.* 2023;12(12):2620. doi:10.3390/electronics12122620.
97. Tang Z, Lu J, Chen Z, Qi F, Zhang L. Improved Pest-YOLO: real-time pest detection based on efficient channel attention mechanism and transformer encoder. *Ecol Inform.* 2023;78:102340. doi:10.1016/j.ecoinf.2023.102340.
98. Venkatasachandran P, Iyapparaja M. GNViT-an enhanced image-based groundnut pest classification using vision transformer (ViT) model. *PLoS One.* 2024;19(3):e0301174. doi:10.1371/journal.pone.0301174.
99. Xie M, Ye N. Multi-scale and multi-factor ViT attention model for classification and detection of pest and disease in agriculture. *Appl Sci.* 2024;14(13):5797. doi:10.3390/app14135797.
100. Li YC, Li R, Zhu JB. Hybrid model based on CNN-transformer for tomato pest and disease identification. *IAENG Int J Comput Sci.* 2025;52(1):223–32.
101. Yang S, Feng Q, Yan W, Zhou W, Yang W. Recognizing few-shot crop diseases using multimodal-guided visual transformer. *Trans Chin Soc Agric Eng.* 2025;41(6):195–203. (In Chinese). doi:10.11975/j.issn.1002-6819.202409189.
102. Yan C, Liang Z, Cheng H, Li S, Yang G, Li Z, et al. CDIP-ChatGLM3: a dual-model approach integrating computer vision and language modeling for crop disease identification and prescription. *Comput Electron Agric.* 2025;236:110442. doi:10.1016/j.compag.2025.110442.
103. Qing J, Deng X, Lan Y, Li Z. GPT-aided diagnosis on agricultural image based on a new light YOLOPC. *Comput Electron Agric.* 2023;213:108168. doi:10.1016/j.compag.2023.108168.
104. Wang Y, Wang F, Chen W, Lv B, Liu M, Kong X, et al. A large language model for multimodal identification of crop diseases and pests. *Sci Rep.* 2025;15(1):21959. doi:10.1038/s41598-025-01908-0.
105. Ahmedm Y, Luan R. AgriM-LLM: an agriculture-specific multimodal large language model for intelligent crop disease and pest management. *Preprints.* 2026. doi:10.20944/preprints202601.0646.v1.
106. Yu K, Ma W, Chen S, Wu X, Wang Y, Zhou T. CPCR: crop pests cross-modal retrieval with prior instruction via multimodal large language model. *Comput Electron Agric.* 2026;245:111546. doi:10.1016/j.compag.2026.111546.

107. Hu T, Li X, Cao K, Qi R, Du J, Dong W, et al. PestCLIP: an incremental pest recognition framework based on a vision-language model. *Pest Manage Sci.* 2026;82(6):5598–611. doi:10.1002/ps.70664.
108. Yuan Z, Liu K, Li S, Peng R, Leybourne D, Musa N, et al. PEZEGO: a precision agriculture system based on large language models and Internet of Things for pest management. *IEEE Internet Things J.* 2025;12(23):49120–30. doi:10.1109/JIOT.2025.3586374.