

Article

Fine-Grained Sex Classification of *Chilo suppressalis* Based on Edge-Cloud Collaboration System

Shengjie Yang ¹, Luyue Wang ¹, Yingchao Zhan ¹, Miao Lu ¹, Yige Zheng ¹, Pan Ma ², Lixing Wei ³, Wen Zhang ^{1,4} 
and Shuangxi Liu ^{1,5,*} 

¹ College of Mechanical and Electronic Engineering, Shandong Agricultural University, Tai'an 271018, China; 2024120823@sdau.edu.cn (S.Y.); 2025110402@sdau.edu.cn (L.W.); 2024120828@sdau.edu.cn (Y.Z.); 2024120799@sdau.edu.cn (M.L.); 2025120742@sdau.edu.cn (Y.Z.); wenzhang098@sdau.edu.cn (W.Z.)

² Jinan Academy of Agricultural Sciences, Jinan 250300, China; 2025120749@sdau.edu.cn

³ Dongying Academy of Agricultural Sciences, Dongying 257091, China; 2023110536@sdau.edu.cn

⁴ Shandong Key Laboratory of Intelligent Production Technology and Equipment for Facility Horticulture, Tai'an 271018, China

⁵ Shandong Engineering Research Center of Agricultural Equipment Intelligentization, Tai'an 271018, China

* Correspondence: 2023110453@sdau.edu.cn; Tel.: +86-15065801249

Abstract

The sex ratio of the rice stem borer, *Chilo suppressalis* (Walker), is important for assessing reproductive potential and outbreak risk, but automated sex classification in field trap images is hindered by small targets, variable postures, subtle morphological differences and complex backgrounds. An edge–cloud collaborative system was developed for fine-grained sex classification of the rice stem borer. The edge terminal performs image acquisition, target detection, ROI extraction, and foreground enhancement, while the cloud platform conducts sex classification and result management. To improve small-target localization, YOLO-CSNet was constructed by integrating a global attention mechanism and adaptive spatial feature fusion into YOLO11n. Within detection-constrained ROIs, Haar-like features and an AdaBoost discriminator were used to suppress tray textures, shadows and non-target regions. For sex classification, DRS-ViT combines convolutional patch embedding, a discriminative-region soft-weighting module and a KANsformer encoder to represent weak sex-related cues and model global structural relationships. YOLO-CSNet achieved precision, recall, mAP@0.5, and mAP@0.5:0.95 values of 94.36%, 93.19%, 95.32%, and 73.95%, respectively. DRS-ViT achieved accuracy, precision, recall, and F1-score values of 93.93%, 94.68%, 93.29%, and 93.91%, respectively. In a temporally independent field deployment conducted at one Shandong site from May to June 2026, the system achieved an image-upload success rate of 91.7% and an end-to-end accuracy of 87.9%. The field results support the feasibility of the workflow under the tested conditions. Future work will extend field validation across multiple sites and seasons to evaluate system generalizability.



Academic Editor: Xanthoula Eirini Pantazi

Received: 21 July 2026

Revised: 31 August 2026

Accepted: 6 September 2026

Published: 8 September 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: rice pest monitoring; small-target detection; fine-grained classification; foreground enhancement; smart trapping

1. Introduction

Rice is one of the most important food crops worldwide, particularly in Asia, and its stable and high yield is essential for ensuring food security [1]. Lepidopteran pests such as the rice stem borer, *Chilo suppressalis* (Walker), pose a long-term threat to rice yield and quality. Their larvae bore into rice stems and damage vascular tissues, which easily

causes dead hearts, white heads, reduced seed-setting rates, and, in severe cases, yield loss and quality deterioration [2]. Over the past decade and more, the area affected by crop diseases and insect pests in China has continued to increase, posing a persistent threat to food security [3]. Field pest occurrence is characterized by sudden outbreaks, regional distribution, and dynamic variation. Traditional methods, including manual field investigation, sorting after trapping, and expert identification, suffer from low efficiency, strong subjectivity, and insufficient capacity for continuous monitoring, making it difficult to meet the demands of modern agriculture for real-time monitoring and rapid early warning [4,5]. Therefore, the accurate acquisition of pest species, population density, and sex ratio is of value for assessing population outbreak risk, formulating precise control strategies, and reducing unnecessary pesticide application [6–8].

With the advancement of machine vision, deep learning, and edge computing technologies, image-based intelligent identification of rice pests and diseases has received increasing attention [9]. Existing research has focused on aspects such as image acquisition, object detection, classification and recognition, network architecture optimization, and prediction and early warning for rice pests and diseases [10]. In recent years, deep visual models, including convolutional neural networks, YOLO-based detectors, and Vision Transformer architectures, have been widely applied to agricultural pest identification, object localization, population counting, and fine-grained recognition tasks [11–13]. In 2025, Yin et al. [14] proposed YOLO-RMD by integrating receptive field attention, mixed local channel attention, and multi-scale detection to improve rice pest detection in complex field environments. The model achieved an mAP@0.5 of 98.2% and demonstrated its feasibility for real-time deployment on edge devices. In 2025, Qiang et al. [15] proposed YOLO-PEST by integrating occlusion augmentation, ConvNeXt-based multi-scale feature fusion, and CoTAttention to enhance the robustness of small-pest detection. In 2025, Zheng et al. [16] proposed Paddy-YOLO and constructed the R2000 dataset to improve rice pest detection under complex backgrounds and high inter-class similarity. In 2026, Yuan et al. [17] developed HDA-YOLO using asymmetric dynamic downsampling to improve the multi-scale discrimination of small and camouflaged rice pests. These strategies improve the detection of small, occluded, camouflaged, or visually similar pests under complex field backgrounds. Smart traps and edge-deployed models have also enabled automated in-trap detection and field-based counting under limited computing and communication resources [18–20]. However, these systems primarily address species-level detection, counting, or selective capture and do not integrate individual localization, foreground optimization, and cloud-based analysis for sex classification of the pest. In 2024, Fu et al. [21] improved the Vision Transformer using image patch partitioning and self-attention to extract discriminative pest features and validated the method in a seven-class recognition task. In 2024, Fang et al. [22] proposed Pest-ConFormer by combining convolutional networks, Transformer, and multi-scale weakly supervised feature selection, achieving an accuracy of 77.81% on the IP102 dataset. In 2025, Kim et al. [23] developed ROI-ViT using soft segmentation, class activation maps, and multi-scale cross-attention, achieving accuracies of 81.81%, 99.64%, and 84.66% on the IP102, D0, and SauTeg datasets, respectively. In 2023, Zhang et al. [24] proposed MMFGT by integrating self-supervised contrastive learning, fine-grained feature attention, and image-text multimodal fusion, achieving an accuracy of 98.12% in few-shot pest recognition. Overall, existing research on visual recognition of rice pests using deep learning has made progress in object detection, species identification, and population counting. Most methods remain focused on the “species-quantity” level and have not yet advanced to identifying sex differences within the same species. The sex ratio of the second-generation rice stem borer directly affects the population’s reproductive potential and the scale of subsequent outbreaks, making it a crucial reference indicator

for pest forecasting and control [25,26]. Field-acquired pest images additionally vary in background, scale, orientation, specimen condition, and image quality, which can create a gap between model-development and deployment data [27,28]. An operational solution must therefore combine reliable individual localization, preservation of weak sex-related cues, and computation suited to field devices.

To address these challenges, this study developed an edge–cloud collaborative workflow for fine-grained sex classification of the rice stem borer. The main contributions are threefold. First, YOLO-CSNet and detection-constrained foreground enhancement are configured to localize small targets and suppress trapping-background interference before classification. Second, the Discriminative-Region Soft-Weighting Module (SWV) integrates multi-head voting for cross-head agreement with continuous attention intensity, spatial smoothing, and residual soft recalibration. The resulting weighting strategy retains information from multiple weak-response tokens rather than selecting only the highest-response regions. Third, detection, ROI cropping, and foreground enhancement are deployed at the edge terminal, whereas DRS-ViT classification is performed in the cloud. The complete workflow is evaluated through a temporally held-out field deployment.

2. Materials and Methods

2.1. System Development

The edge–cloud collaborative fine-grained sex classification system is based on a portable push-pull image acquisition device for rice pests and integrates edge-side object detection and candidate-region foreground enhancement with cloud-based fine-grained sex classification and remote data management. It implements a complete workflow covering pest trapping, image acquisition, target localization, ROI optimization, sex classification, and result uploading, as shown in Figure 1.

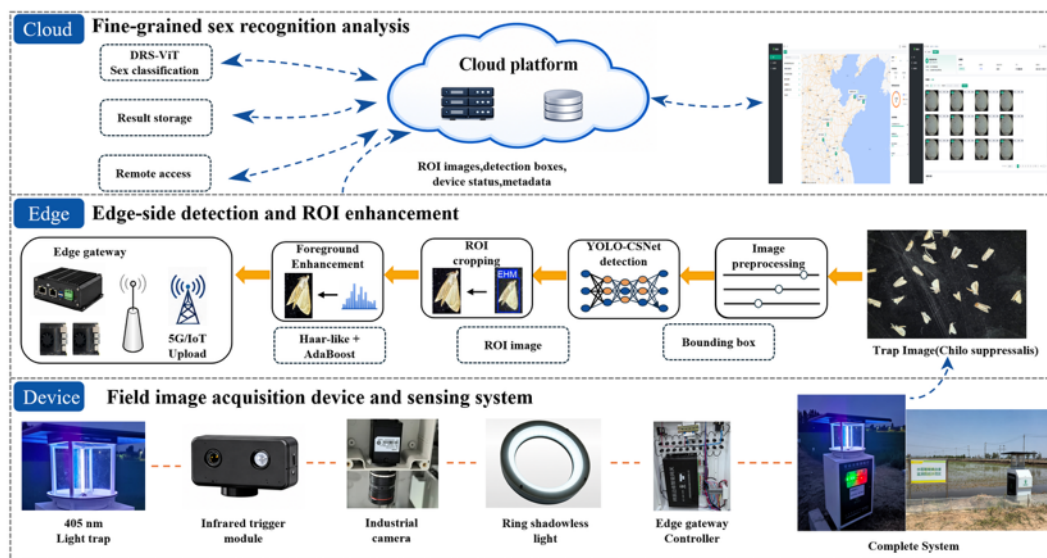


Figure 1. Edge–cloud collaborative fine-grained sex classification system for the rice stem borer.

The edge–cloud task allocation was determined by the processing requirements of the detection and classification stages. Target detection, ROI cropping, and foreground enhancement are performed at the edge terminal to reduce non-target image content and data transmission volume. Fine-grained classification is performed on the cloud server because DRS-ViT requires greater computational capacity to model subtle local features and their global morphological relationships. This configuration addresses both the resource constraints of field devices and the weak sex-related differences in rice stem borer images.

The hardware configuration described in this section corresponds to the improved device used for the 2026 field validation. This device was developed by modifying the original pest-monitoring device supplied by Shandong Xiangchen Technology Group Co., Ltd. (Jinan, China) In contrast, the 2023–2025 model-development images were acquired using the original device before these modifications. An annular shadowless light source was used to reduce uneven illumination and insect shadows. Highly light-absorptive materials were applied to the inner wall and insect-collecting turntable to reduce specular reflections and background interference. The push-pull structure removes inactive insects from the imaging area, thereby limiting specimen accumulation and repeated image acquisition. These measures are intended to improve the consistency and clarity of the acquired images before object detection and sex classification.

During device operation, a 405 nm short-wavelength insect-attracting light source is used to lure adult rice stem borers [29]. After a pest enters the imaging area, the infrared detection module triggers the industrial camera to acquire an image. The image is transmitted to the edge terminal, where YOLO-CSNet localizes the rice stem borer targets and generates candidate ROIs. Foreground enhancement is then performed within each candidate ROI to reduce tray textures, local shadows, and non-target regions. The processed ROI images, detection boxes, and device-status information are uploaded through a 5G communication gateway. The cloud server applies DRS-ViT for sex classification and stores the results for remote access and sex-structure analysis. By transmitting candidate ROI images rather than the original full images, the system reduces the amount of image data uploaded to the cloud. The device information is shown in Table 1.

Table 1. Equipment selection and performance parameters.

Device	Model	Function
Industrial camera	Hikvision MV-CU120-10GC	Pest image acquisition
Infrared sensor	TSMP95000TR	Imaging trigger
5G edge gateway control unit	Xiang Chen Intelligent Gateway XC-GW-CBD	Edge computing, 5G communication, local storage
Voltage regulator chip	AMS1117-3.3	3.3 V regulated power supply
USB-to-serial converter chip	CH340C	USB-to-serial communication and program downloading

All equipment used in this study was supplied by Shandong Xiangchen Technology Group Co., Ltd. (Jinan, China).

2.2. Pest Image Acquisition and Dataset Construction

Rice stem borer images were acquired at three experimental sites in Shandong Province, including Jinan Wuzhou Urban Agriculture Co., Ltd., the pest monitoring site of the Shandong Academy of Agricultural Sciences, and the rice base of Jining Lüjian Agricultural Technology Co., Ltd. Images acquired at these sites from 2023 to 2025 were collected using the original trapping device before the hardware modifications described in Section 2.1, as shown in Figure 2a. These images covered different field environments and constituted the model-development dataset used for model training, validation, and internal testing. Model training, hyperparameter optimization, threshold selection, and model selection were conducted from September 2025 to March 2026. Subsequently, the improved trapping device described in Section 2.1, shown in Figure 2b, was used to collect a temporally held-out field-validation dataset at Jinan Wuzhou Urban Agriculture Co., Ltd. (36.68° N,

116.95° E) from May to June 2026. None of the 2026 field-validation images were used during model development. All field images were uniformly resized to 640×640 pixels.

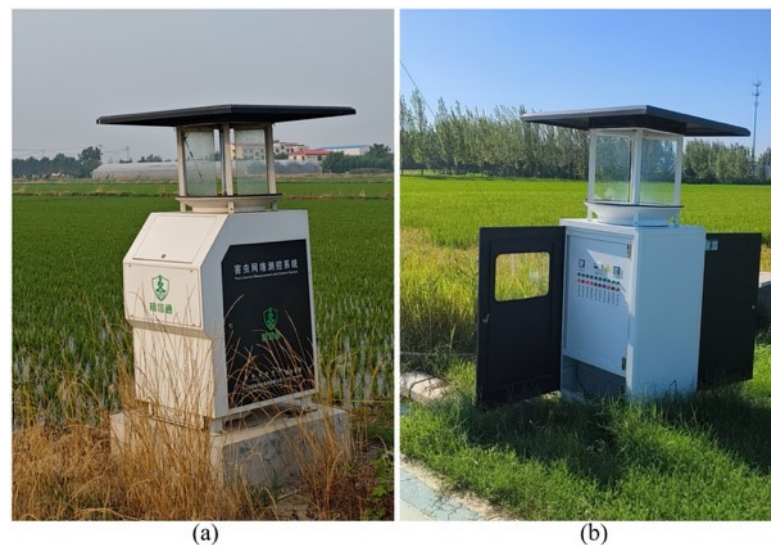


Figure 2. Image acquisition devices used during the two study periods. (a) Original trapping device used to acquire the model-development images from 2023 to 2025. The Chinese inscriptions identify the unit as a “Pest Internet Measurement and Control System” and display the “Zhixintong” brand name (literally, “Plant Information Link”). (b) Improved trapping device used to acquire the temporally held-out field-validation images from May to June 2026. Both devices were manufactured by Shandong Xiangchen Technology Group Co., Ltd. (Jinan, China).

During 2023–2025, inactive insects captured at the experimental sites were transported to the laboratory for taxonomic examination. Species identities were initially assigned by the research team using established morphological criteria and subsequently verified by entomology experts. Following verification, 2566 multi-insect trap images acquired periodically across sites and sampling periods were selected when target instances could be reliably identified and annotated. Expert-confirmed rice stem borer and rice leaffolder instances were annotated as separate classes for detector development. All expert-confirmed rice stem borer instances were annotated with class-specific bounding boxes using LabelImg (version 1.8.6) to construct the Rice Insect dataset for object detection. Before augmentation, the original images were divided into training, validation, and test sets at a ratio of 8:1:1. Data augmentation, including horizontal flipping, random rotation, brightness adjustment, Gaussian noise, random occlusion, and random cropping, was applied only to the training set. Representative outputs from these augmentation operations are shown in Figure 3. After augmentation, the Rice Insect dataset contained 8211 images and was used to train YOLO-CSNet for target detection and localization under complex trapping backgrounds.

Trap-caught inactive adults were first identified as the rice stem borer by the research team using species-specific morphological characteristics. Initial sex labels were then assigned according to established diagnostic features, including body size, forewing coloration and markings, and abdominal morphology, following relevant morphological references and identification criteria [30–32]. Before dataset partitioning, the retained specimens and their sex labels were independently reviewed by entomology experts from the College of Plant Protection, Shandong Agricultural University who had not participated in the initial annotation. Specimens with uncertain species or sex assignments, or with unresolved disagreement between the initial annotation and expert review, were excluded. Only expert-confirmed specimens and sex labels were used as ground truth for model training and evaluation.

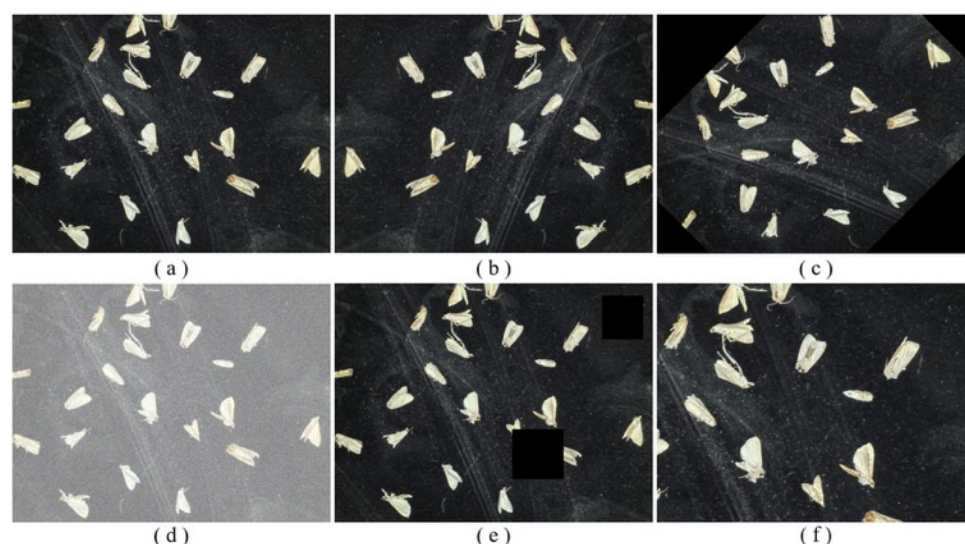


Figure 3. Examples of data augmentation. (a) Original image. (b) Horizontal flipping. (c) Random rotation. (d) Gaussian noise. (e) Random occlusion. (f) Random cropping.

The confirmed specimens were photographed under standardized conditions to construct the controlled sex-classification dataset for DRS-ViT. Camera position, imaging distance, exposure settings, background, and shadowless illumination were fixed to reduce image variation unrelated to sex. Specimens were assigned to predefined viewing orientations to capture variation in wing, abdominal, head, and body-axis morphology relevant to sex identification, as shown in Figure 4. The dataset comprised 1020 single-insect images from 1020 independent specimens, including 498 females and 522 males. Images were stratified at the specimen level while preserving the female-to-male ratio, yielding 714 training images (348 female and 366 male), 153 validation images (75 female and 78 male), and 153 test images (75 female and 78 male). Data augmentation was applied only to the training set; the validation and test images remained unaugmented.

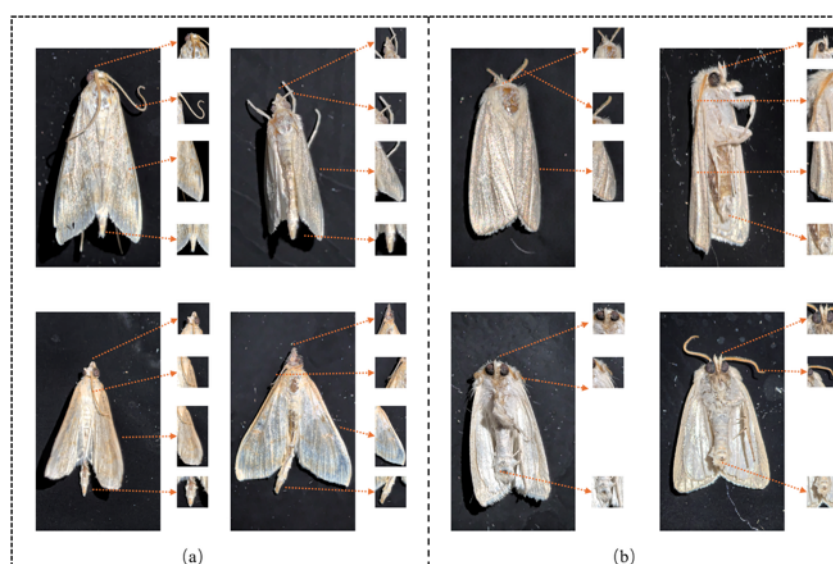


Figure 4. Representative images and selected morphological regions from the controlled sex-classification dataset. (a) Male *Chilo suppressalis* specimens. (b) Female *Chilo suppressalis* specimens.

2.3. YOLO-CSNet Architecture

YOLO11n was selected as the edge-side baseline because of its compact architecture and inference speed [33]. YOLO-CSNet incorporates the Global Attention Mechanism

(GAM) [34] to recalibrate channel and spatial responses and an ASFF detection head to align and fuse multiscale features. The resulting architecture was designed to strengthen small-target representation while retaining a compact resource footprint. The structure of YOLO-CSNet is shown in Figure 5.

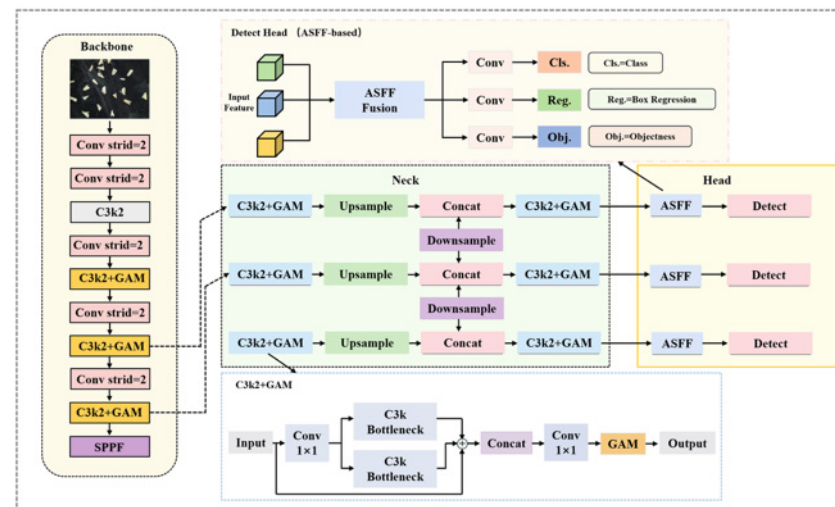


Figure 5. YOLO-CSNet architecture diagram.

2.3.1. GAM Attention Mechanism

To recalibrate channel-wise and spatial feature responses, this study introduced the Global Attention Mechanism (GAM) into YOLO11n. GAM sequentially applies channel attention and spatial attention to the input feature map, as shown in Figure 6.

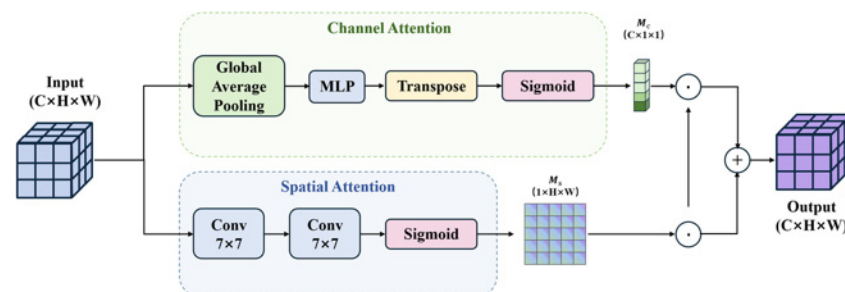


Figure 6. GAM structure diagram.

Let the input feature map be denoted as F . The enhanced feature after the channel attention branch can be expressed as:

$$F_c = M_c(F) \odot F \quad (1)$$

$$F_s = M_s(F) \odot F \quad (2)$$

where $M_c(F)$ and $M_s(F)$ denote the channel attention weights and spatial attention weights, respectively, and $\odot$ denotes element-wise multiplication. Subsequently, the enhanced features produced by the two branches are fused, and the final output of the GAM module can be written as:

$$F_{(out)} = F_c + F_s = M_c(F) \odot F + M_s(F) \odot F \quad (3)$$

2.3.2. Optimization of the ASFF Module

To achieve adaptive multiscale feature fusion, this study introduces an ASFF-based detection head into YOLO11n. This module aligns the resolution and channel dimensions

of three neck outputs and assigns spatially adaptive fusion weights. The fused features are used for classification and bounding-box regression, as shown in Figure 7.

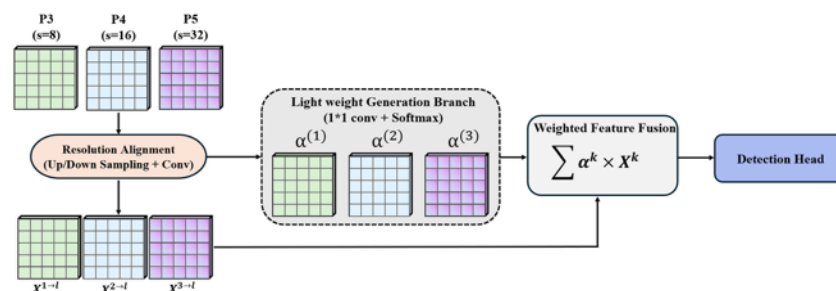


Figure 7. ASFF structure diagram.

Let the target fused feature at the l -th detection scale be denoted as $Y^{(l)}$, and its inputs are derived from feature maps at three scales:

$$X^{(k)} \in \mathbb{R}^{H_k \times W_k \times C_k}, k \in \{1, 2, 3\} \quad (4)$$

where $k \in (1, 2, 3)$ corresponds to shallow, intermediate, and deep features, respectively.

During feature alignment, upsampling adopts interpolation and 1×1 convolution to adjust the resolution and the number of channels, while downsampling uses 3×3 convolution with a stride of 2 to synchronously adjust the spatial size and channel dimension. Let U and D denote the upsampling and downsampling operations, respectively:

$$X^{(k \rightarrow l)} = \begin{cases} U(X^k), & k < l \\ D(X^k), & k > l \\ X^k, & k = l \end{cases} \quad (5)$$

For each spatial location (i, j) , learnable weights $\alpha_k^{(l)}(i, j)$ are assigned to features at different scales, satisfying:

$$\sum_{k=1}^3 \alpha_k^{(l)}(i, j) = 1, \alpha_k^{(l)}(i, j) \in [0, 1] \quad (6)$$

The above weights are obtained by applying Softmax normalization to the scale response scores predicted by the network:

$$\alpha_k^{(l)}(i, j) = \frac{\exp(w_k^{(l)}(i, j))}{\sum_{m=1}^3 \exp(w_m^{(l)}(i, j))} \quad (7)$$

where $w_m^{(l)}(i, j)$ denotes the scale response score generated by the lightweight convolutional weighting branch.

Finally, the fused feature at the l -th detection scale is represented as:

$$Y^{(l)}(i, j) = \sum_{k=1}^3 \alpha_k^{(l)}(i, j) \cdot X^{(k \rightarrow l)}(i, j) \quad (8)$$

Unlike conventional fixed feature fusion methods, the improved detection head adaptively adjusts the fusion ratio between shallow detail features and deep semantic features according to the scale variation and spatial distribution of pests in trap images. The aligned

feature maps are combined using the spatially adaptive weights defined in Equation (8), and the resulting features are used for classification and bounding-box regression.

2.4. Detection ROI-Based Foreground Enhancement

The detection boxes output by YOLO-CSNet can localize individual rice stem borers, but the resulting candidate ROIs may still contain background interference, including insect-collecting tray textures, local shadows, device reflections, and other non-target content. Therefore, ROI foreground enhancement is used as a task-specific preprocessing step to obtain a background-suppressed single-insect image from each detection ROI for subsequent cloud-based sex classification. This stage is intended to reduce non-target content rather than to produce pixel-accurate semantic masks. Within each detection ROI, Haar-like features and AdaBoost are used to distinguish the insect foreground from the residual background and generate background-suppressed images without introducing a separate deep segmentation network into the edge-side pipeline.

During processing, rice stem borer ROI images are cropped according to the detection boxes. The main insect body regions are used as positive samples, while the insect-collecting tray background, shadows, reflections, and other non-insect regions are used as negative samples. Training samples are constructed with a positive-to-negative sample ratio of 1:5. Based on these samples, multi-scale Haar-like features are extracted to describe morphological information, including insect body edges, the body axis, and local textures, and the AdaBoost algorithm is used to train a foreground–background classifier. After 100 iterations, an insect body region discrimination model is obtained and applied to local discrimination and foreground enhancement of candidate ROIs, generating background-suppressed insect images, as shown in Figure 8.

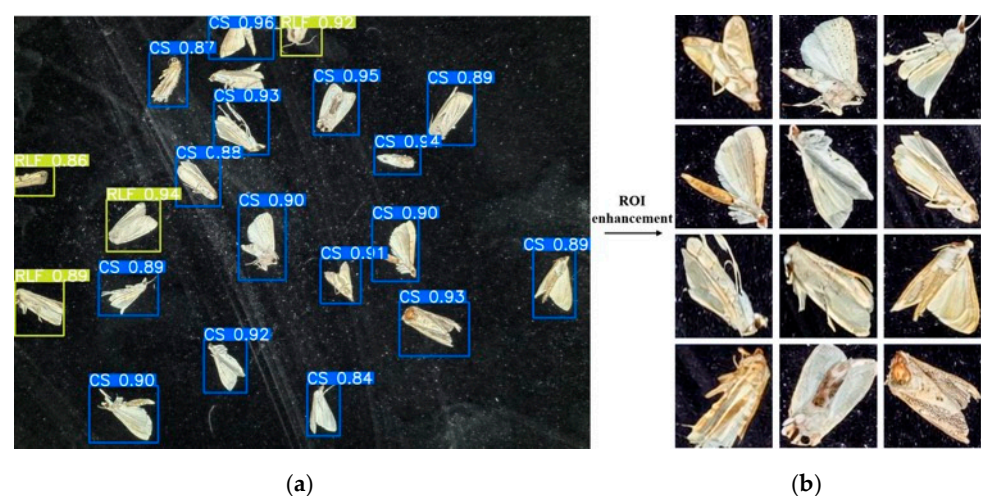


Figure 8. Results of ROI-based foreground extraction. (a) YOLO-CSNet detection results. Blue and yellow bounding boxes indicate *Chilo suppressalis* (CS) and the rice leaf folder, *Cnaphalocrocis medinalis* (RLF), respectively. The numeric values adjacent to the class labels are the corresponding detection confidence scores. (b) Representative regions of interest after foreground enhancement. ROI, region of interest.

2.5. DRS-ViT Architecture

After ROI extraction and insect foreground enhancement, sex-related differences between female and male rice stem borers remained subtle, particularly in local morphological structures and textures. To address this problem, this study constructed DRS-ViT by integrating convolutional patch embedding (CPE), a discriminative-region soft-weighting module, and a KANsformer encoder. CPE extracts local features and maps them into token sequences, while the soft-weighting module recalibrates local tokens. The KANsformer

encoder further models global relationships among these tokens. The structure of DRS-ViT is shown in Figure 9.

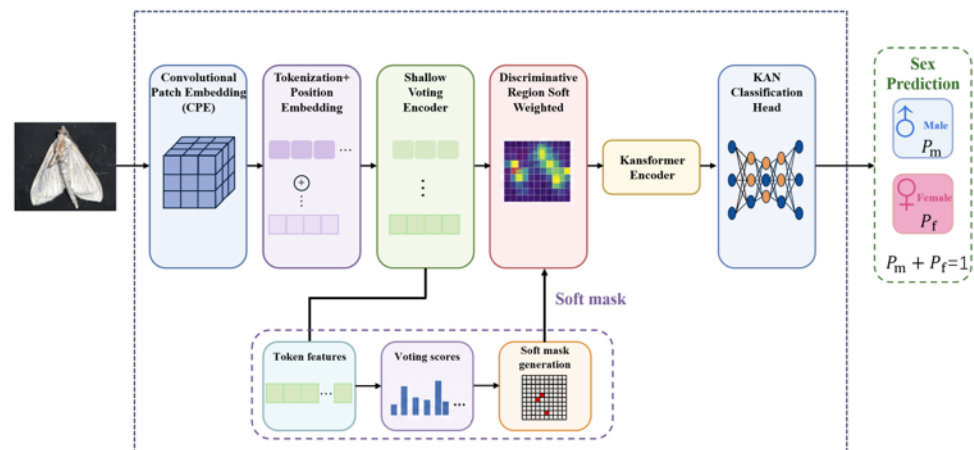


Figure 9. DRS-ViT network structure diagram.

2.5.1. CNN-Based Local Feature Extraction and Tokenized Representation

After ROI extraction and foreground enhancement, morphological differences between expert-confirmed female and male specimens remained visually subtle. Standard Vision Transformer patch projection can disrupt local structural continuity and weaken local-feature representation. To address this issue, DRS-ViT replaces linear patch projection with a lightweight convolutional patch-embedding (CPE) module, as shown in Figure 10. Inputs were resized to 224×224 pixels and processed by two 3×3 convolutional layers with a stride of 2 and padding of 1, followed by ReLU activation and 2×2 pooling. The resulting 28×28 feature map was flattened into 784 local tokens, linearly projected, prepended with a class token, and combined with learnable positional embeddings.



Figure 10. CPE structure diagram.

The cloud-side classification input was resized to 224×224 . The two stride = 2 convolutional layers and the 2×2 pooling layer produced a 28×28 feature map, which was flattened into $N = 784$ local tokens.

2.5.2. Discriminative-Region Soft-Weighting Module

Before the main encoder, this study introduced the Discriminative-Region Soft-Weighting (SWV) module to estimate the importance of local tokens. A shallow attention block generates class-token-to-local-token responses, which are combined through multi-head voting and mean attention. The resulting response map is spatially smoothed, converted into continuous weights, and applied to the local tokens through residual recalibration. In this study, “discriminative region” refers operationally to image-token locations receiving relatively high class-to-token responses. SWV was trained using image-level sex labels without anatomical-region supervision. The workflow is shown in Figure 11.

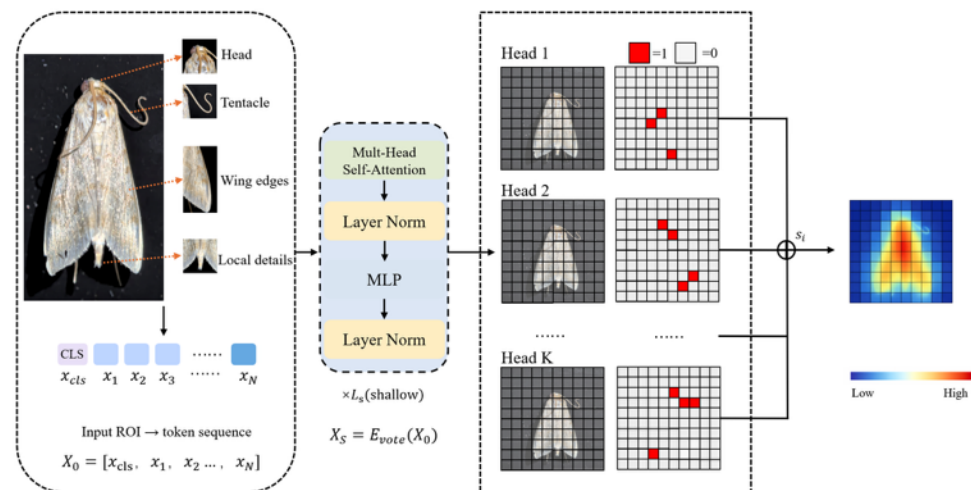


Figure 11. Structure of the shallow multi-head voting-guided discriminative-region soft-weighting module.

Unlike binary token masking, SWV combines multi-head voting with continuous attention intensity. Multi-head voting captures agreement across attention heads, whereas continuous responses retain information from weakly activated regions. Spatial smoothing incorporates local consistency, and residual soft recalibration preserves the original feature baseline. This design is intended for sex-related cues that are distributed across multiple weak-response regions rather than concentrated in one dominant region.

Let the feature map input sequence be denoted as X_0 :

$$X_0 = [x_{cls}, x_1, x_2, \dots, x_N] \in \mathbb{R}^{(N+1) \times d} \quad (9)$$

where x_{cls} denotes the class token, x_i denotes the i -th local token, N denotes the total number of local tokens, and d denotes the embedding dimension. First, X_0 is input into the shallow voting encoder $E_{vote}(x)$, which consists of a lightweight Transformer block, namely $L_s = 1$, to obtain the shallow contextual representation:

$$X_s = E_{vote}(X_0), \quad X_s \in \mathbb{R}^{(N+1) \times d} \quad (10)$$

The shallow voting encoder contains one lightweight Transformer block and produces the contextual representation defined in Equation (10).

The last multi-head self-attention layer of the shallow encoder contains $K = 12$ attention heads. For the k -th head, the input features are linearly transformed and mapped into the Query, Key, and Value vector spaces:

$$Q^{(k)} = X_s W_Q^{(k)}, \quad K^{(k)} = X_s W_K^{(k)}, \quad V^{(k)} = X_s W_V^{(k)} \quad (11)$$

where $W_Q^{(k)}, W_K^{(k)}, W_V^{(k)} \in \mathbb{R}^{d \times d_h}$, $d_h = d/K$ denotes the feature dimension of a single head. Subsequently, the attention matrix of the k -th head can be expressed as:

$$A^{(k)} = \text{softmax} \left(\frac{Q^{(k)} (K^{(k)})^T}{\sqrt{d_h}} \right), \quad A^{(k)} \in \mathbb{R}^{(N+1) \times (N+1)} \quad (12)$$

Considering the class token aggregates information across the image, its responses to local tokens were used as the basis for token-importance estimation. For the k -th attention head, the response vector from the class token to the local tokens is extracted as:

$$a^{(k)} = A_{0,1:N}^{(k)} \in \mathbb{R}^N \quad (13)$$

Class-to-token responses were extracted from each of the 12 attention heads using Equation (13). For each head, the 100 highest-response tokens, corresponding to 12.8% of the 784 local tokens, formed the binary voting vector in Equation (14). Votes were summed across heads using Equation (15), while mean attention was retained as the continuous response term in Equation (16). The two terms were combined in Equation (17), with α fixed at 0.6 [35] to assign greater weight to cross-head agreement while retaining continuous responses.

$$b_i^{(k)} = \begin{cases} 1, & a_i^{(k)} \in \text{Top-}r(a^{(k)}), \quad i = 1, 2, \dots, N \\ 0, & \text{otherwise} \end{cases} \quad (14)$$

$$m_i = \sum_{k=1}^K b_i^{(k)}, \quad i = 1, 2, \dots, N \quad (15)$$

$$\bar{a}_i = \frac{1}{K} \sum_{k=1}^K a_i^{(k)} \quad (16)$$

$$s_i = \alpha \frac{m_i}{K} + (1 - \alpha) \bar{a}_i, \quad 0 \leq \alpha \leq 1 \quad (17)$$

The score vector was reshaped according to the original token grid and smoothed using the fixed Gaussian convolution defined in Equation (18). This operation preserved token order and reduced isolated responses.

$$S \in \mathbb{R}^{H' \times W'}, S' = \text{Conv}_{fix}(S) \quad (18)$$

The smoothed map was flattened and converted into continuous weights using Equation (19). Its median was used as μ , and T was fixed at 0.2 [36]. The weights were applied through residual recalibration in Equation (20), with λ fixed at 0.5 [37], limiting the maximum amplification factor to 1.5. The class token was not rescaled, and the final sequence is defined in Equation (21).

$$w_i = \sigma\left(\frac{s'_i - \mu}{T}\right), \quad i = 1, 2, \dots, N \quad (19)$$

$$\tilde{x}_i = (1 + \lambda w_i) x_i^s, \quad i = 1, 2, \dots, N \quad (20)$$

$$\tilde{X} = [x_{cls}^s, \tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_N] \in \mathbb{R}^{(N+1) \times d} \quad (21)$$

2.5.3. KANsformer Encoder

To model nonlinear interactions among local token features, this study constructed a KANsformer encoder by replacing the MLP in each Transformer block with a KAN-based nonlinear mapping. The encoder retains multi-head self-attention, layer normalization, and residual connections. KAN places learnable univariate nonlinear functions on the network edges. Its structure is shown in Figure 12, and its expression is given as:

$$f(x) = f(x_1, \dots, x_n) = \sum_{q=1}^{2n+1} \Phi_q \left(\sum_{p=1}^n \phi_{q,p}(x_p) \right) \quad (22)$$

where x_p denotes the p -th component of the input vector x , $\phi_{q,p}(x)$ denotes a learnable univariate continuous function, and $\Phi_q(x)$ denotes the outer composition function.

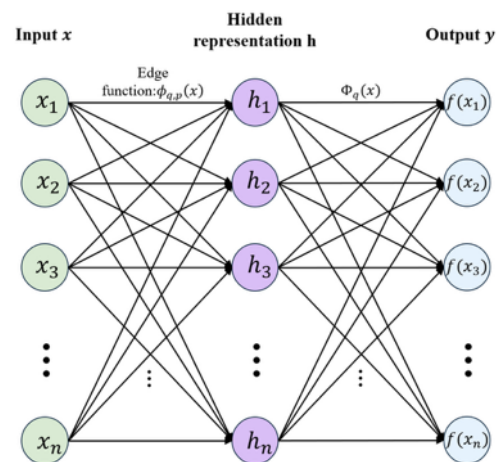


Figure 12. Structure diagram of the KAN function.

Compared with a traditional MLP, KAN helps enhance the modeling capability for nonlinear interactions. This enables the KANsformer encoder to achieve stronger stability and adaptability when handling contour deformation, complex wing margins, and diverse local textures, thereby improving the discrimination accuracy of male–female differences.

The improved KANsformer encoder in this study consists of L_k stacked encoding blocks. Each encoding block includes a multi-head attention sublayer and a KAN-based nonlinear mapping sublayer, with normalization and residual connections applied around each sublayer. The computational process can be expressed as:

$$\hat{X}^{(l)} = X^{(l-1)} + MHSA\left(LN\left(X^{(l-1)}\right)\right) \quad (23)$$

$$X^{(l)} = \hat{X}^{(l)} + \kappa\left(LN\left(\hat{X}^{(l)}\right)\right), l = 1, 2, \dots, L_k \quad (24)$$

where $LN(x)$ denotes layer normalization, $MHSA(x)$ denotes the multi-head self-attention operator, and $\kappa(x)$ denotes the KAN-based nonlinear mapping sublayer. This structure preserves the ability of the Transformer to model long-range dependencies among tokens, while KAN strengthens the modeling of nonlinear interactions among fine-grained features from multiple body regions. The global semantic representation output by the encoder is transformed into female and male class probabilities through a KAN classification head and Softmax, and cross-entropy loss is used for supervised optimization during training:

$$\mathcal{L}_{cls} = - \sum_{c=1}^2 y_c \log \hat{y}_c \quad (25)$$

where y_c denotes the ground-truth class label, and $\hat{y}_c$ denotes the predicted probability of the c -th class output by the model. By minimizing this loss function, the model can jointly optimize the processes of front-end local feature extraction, discriminative-region recalibration, and back-end global relationship modeling and classification decision, thereby improving the overall accuracy and robustness of rice stem borer sex recognition.

2.6. Experimental Setup and Hyperparameter Settings

The experimental environment was configured on a server running the Windows 11 Professional operating system. The server was equipped with an Intel(R) Core(TM) i9-14900K processor and an NVIDIA 4090 graphics processor. CUDA 11.8 was installed,

and the Python programs were developed using Python 3.9 based on the PyTorch 2.0.0 deep learning framework. The hardware configuration and hyperparameter settings are listed in Table 2.

Table 2. Experimental environment and hyperparameter settings.

Category	Parameter	YOLO-CSNet	DRS-ViT
Experimental environment	Operating system	Windows11 Professional	Windows11 Professional
	Development environment	Anaconda	Anaconda
	Programming language	Python3.9	Python3.9
	Deep learning framework	PyTorch2.0.0	PyTorch2.0.0
	CPU	Intel(R) Core (TM)i9-14900K@3.0 GHz	Intel(R) Core (TM)i9-14900K@3.0 GHz
	GPU	NVIDIA GeForce RTX 4090	NVIDIA GeForce RTX 4090
	CUDA version	CUDA 11.8	CUDA 11.8
	GPU memory	24 GB	24 GB
	RAM	128 GB	128 GB
	Storage	1 TB SSD	1 TB SSD
Training settings	Task	Object detection	Fine-grained sex classification
	Input image size	640 × 640	224 × 224
	Batch size	16	16
	Epochs	300	300
	Optimizer	SGD	SGD
	Initial learning rate	0.001	0.001
	Random seeds	42,3407,2026	42,3407,2026
	Momentum	0.937	0.9
	Weight decay	5 × 10 ^{−4}	5 × 10 ^{−5}

2.7. Field System Validation

Field validation was conducted at Jinan Wuzhou Urban Agriculture Co., Ltd. (36.68° N, 116.95° E) from May to June 2026 using the improved portable push-pull trapping device. Ten experimental units were established, each covering approximately 0.67 ha, and one device was deployed in each unit. All field-validation images were acquired after model development and were excluded from training, hyperparameter optimization, threshold selection, and model selection. The field dataset was therefore temporally independent of model development, although the deployment site had contributed historical images to the 2023–2025 dataset.

A 405 nm light source was used during field deployment [29], and other phototactic insects were occasionally captured. YOLO-CSNet served as the species-filtering stage, and only ROIs assigned to rice stem borer were forwarded to DRS-ViT for sex classification. Other insect taxa were not represented as explicit detector classes. Consequently, morphologically similar non-target insects could be assigned to one of the trained classes and propagate as false-positive detections.

End-to-end evaluation treated edge-side object detection, ROI cropping, insect foreground enhancement, and cloud-based sex classification as a complete recognition workflow. An expert-confirmed rice stem borer individual was counted as correctly recognized only when the IoU between the predicted and ground-truth bounding boxes was at least 0.7 and the predicted sex matched the expert-confirmed label. Ground-truth bounding boxes were manually annotated, whereas sex labels were independently assigned by entomology experts from the College of Plant Protection, Shandong Agricultural University, without access to DRS-ViT predictions. Only individuals with expert-confirmed sex labels were included in the evaluation. System-level accuracy was calculated as the number of individ-

uals satisfying both conditions divided by the total number of evaluated individuals. The internal structure of the device and its field deployment are shown in Figure 13.



Figure 13. Test site and installed equipment. The Chinese inscription on the front panel identifies the unit as an “Intelligent Insect Monitoring Lamp”. The warning label reads “Notice: Experimental equipment, please keep away.” The device was manufactured by Shandong Xiangchen Technology Group Co., Ltd. (Jinan, China).

2.8. Evaluation Metrics

In the object detection stage, precision (P), recall (R), mAP@0.5, and mAP@0.5:0.95 are used to evaluate model performance. Their definitions are shown in Equations (26)–(29):

$$P = \frac{TP}{TP + FP} \quad (26)$$

$$R = \frac{TP}{TP + FN} \quad (27)$$

$$A_P = \int_0^1 P(X) dx \quad (28)$$

$$mAP@0.5 = \frac{1}{m} \sum_i^m A_P(i) \quad (29)$$

where TP denotes the number of correctly identified positive samples, FP denotes the number of samples incorrectly labeled as positive, and FN denotes the number of positive samples that are not detected.

In the sex-classification stage, accuracy, precision, recall, and F1-score were reported. Female samples were defined as the positive class and male samples as the negative class. Accordingly, TP and TN represented correctly classified female and male samples, whereas FP and FN represented male-to-female and female-to-male errors, respectively. Sex-specific recall and row-normalized confusion matrices were additionally used to examine class-specific performance. Their definitions of accuracy and F1-score are shown in Equations (30) and (31):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (30)$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (31)$$

3. Experimental Results

3.1. Training Results of the YOLO-CSNet Model

3.1.1. Model Comparison Experiment

To evaluate the performance of YOLO-CSNet in the rice stem borer detection task, this study selected mainstream object detection models, including Faster R-CNN [38], YOLOv5, YOLOv7, YOLOv8, YOLOv9, YOLOv10, YOLO26n, and the original YOLO11n, as comparative methods and conducted experimental comparisons with the proposed model. To ensure comparability of the results, all models were trained on the same Rice Insect dataset using consistent training parameters for 300 epochs. The evaluation metrics included precision, recall, mAP@0.5, and mAP@0.5:0.95. The experimental results are presented in Table 3, and the performance comparison among different models is shown in Figure 14.

Table 3. YOLO-CSNet comparison test table.

Model	P/%	R/%	mAP@0.5/%	mAP@0.5:0.95/%	FPS
Faster R-CNN	81.38 ± 0.24	80.55 ± 0.23	82.16 ± 0.18	60.59 ± 0.16	23
YOLOv5	83.26 ± 0.23	82.01 ± 0.22	85.47 ± 0.23	66.20 ± 0.14	145
YOLOv7	79.22 ± 0.21	80.50 ± 0.22	83.09 ± 0.20	64.03 ± 0.19	187
YOLOv8n	86.31 ± 0.23	84.26 ± 0.30	86.20 ± 0.14	67.23 ± 0.20	210
YOLOv9t	89.94 ± 0.19	89.86 ± 0.26	90.72 ± 0.12	71.46 ± 0.24	158
YOLOv10n	91.12 ± 0.18	91.30 ± 0.16	91.93 ± 0.19	70.76 ± 0.18	135
YOLOv11n	88.92 ± 0.32	90.65 ± 0.18	90.52 ± 0.23	70.05 ± 0.26	128
YOLO 26n	93.48 ± 0.15	92.77 ± 0.21	92.49 ± 0.18	73.43 ± 0.16	146
YOLO-CSNet	94.36 ± 0.20	93.19 ± 0.23	95.32 ± 0.09	73.95 ± 0.12	117

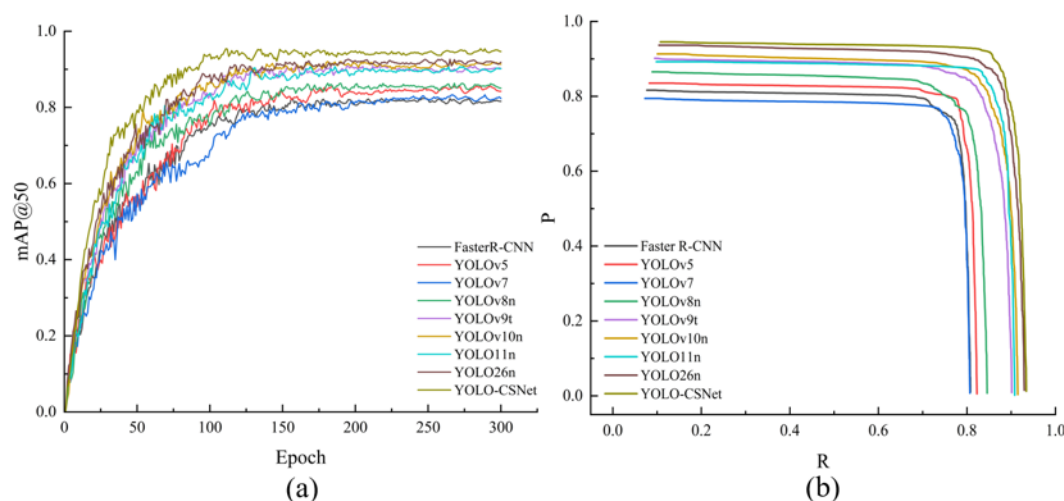


Figure 14. Comparison curves with other mainstream models. (a) mAP@0.5 curves of different models. (b) P–R curves of different models.

As shown in Table 3, YOLO-CSNet achieved the highest mean detection accuracy among the evaluated models, with an mAP@0.5 and mAP@0.5:0.95 values of $95.32 \pm 0.09\%$, and $73.95 \pm 0.12\%$. Compared with the original YOLO11n, YOLO-CSNet improved both detection precision and recall, indicating that the introduced structural optimization strategies enhance the feature representation capability of the model for rice stem borer target regions. Under conditions involving complex trapping backgrounds, variations in insect body scale, and indistinct target-edge features, YOLO-CSNet localized target regions more stably and reduced missed detections and false detections to some extent.

3.1.2. YOLO-CSNet Ablation Study

Table 4 summarizes the ablation results. GAM alone yielded mean precision, recall, and mAP@0.5 values of $90.48 \pm 0.24\%$, $90.44 \pm 0.17\%$, and $91.81 \pm 0.14\%$, respectively. ASFF alone increased these metrics by 3.02, 0.61, and 3.02 percentage points over YOLO11n. Combining GAM and ASFF produced $94.36 \pm 0.20\%$, $93.19 \pm 0.23\%$, and $95.32 \pm 0.09\%$, respectively, the highest values among the evaluated variants.

Table 4. Results of YOLO-CSNet ablation experiments.

NO.	GAM	ASFF	Params/M	P/%	R/%	mAP@0.5/%	FLOPs/G	Model Size/MB
Baseline	×	×	2.6	88.92 ± 0.32	90.65 ± 0.18	90.52 ± 0.23	6.6	5.35
1	✓	×	2.2	90.48 ± 0.24	90.44 ± 0.17	91.81 ± 0.14	6.1	4.6
2	×	✓	2.6	91.94 ± 0.22	91.26 ± 0.24	93.54 ± 0.19	6.4	5.4
3	✓	✓	2.3	94.36 ± 0.20	93.19 ± 0.23	95.32 ± 0.09	6.2	4.8

“×” indicates that the corresponding module was not included, whereas “✓” indicates that the corresponding module was included.

Although YOLO-CSNet maintained a relatively small resource footprint, its inference speed decreased from 128 FPS for YOLO11n to 117 FPS, corresponding to an 8.6% reduction. As shown in Table 4, YOLO-CSNet reduced the parameter count from 2.6 M to 2.3 M, FLOPs from 6.6 G to 6.2 G, and model size from 5.35 MB to 4.80 MB while increasing mAP@0.5 by 4.80 percentage points. Despite the reduced model complexity, the lower FPS may be associated with the additional feature transformations introduced by GAM and ASFF, including channel and spatial attention, multiscale feature alignment, resampling, and spatially adaptive weighting. These operations may introduce memory-access and parallelization overheads that are not fully captured by the parameter count or FLOPs. Therefore, the term “lightweight” in this study refers to the reduced parameter count, FLOPs, and model size rather than to increased raw inference throughput. YOLO-CSNet thus achieves higher detection performance at the cost of a moderate reduction in FPS. FPS was measured under the common environment specified in Table 2 and was used only for relative model comparison. Field-deployment feasibility was assessed separately using the edge-side processing time in Section 3.3.

3.2. Training Results of the DRS-ViT Model

3.2.1. Model Comparison Experiment

DRS-ViT was compared with ResNet50, MobileNetV3-Small, EfficientNet-B0, ConvNeXt-T, ViT-B-16, Swin-T, and TransFG using the controlled sex-classification dataset and the common training protocol described in Sections 2.2 and 2.6. Results are reported in Table 5. DRS-ViT achieved the highest mean values for all four classification metrics among the evaluated models, including a mean accuracy of 93.93%.

As shown in Table 5, DRS-ViT achieved the highest mean accuracy, precision, recall, and F1-score values of $93.93 \pm 0.21\%$, $94.68 \pm 0.17\%$, $93.29 \pm 0.12\%$, and $93.91 \pm 0.19\%$, respectively, and its overall performance is superior to that of all comparative models. Compared with ViT-B-16, DRS-ViT increased mean accuracy and F1-score by 2.90 and 4.69 percentage points, respectively. Compared with TransFG, the corresponding gains were 1.78 and 1.93 percentage points. These results indicate that DRS-ViT improved fine-grained sex-classification performance under the evaluated dataset and training conditions.

Table 5. DRS-ViT comparative test table.

Model	Backbone	Acc/%	P/%	R/%	F1/%	Params/M	Train Time/h
ResNet50	CNN	82.53 $\pm$ 0.14	82.41 $\pm$ 0.30	81.14 $\pm$ 0.23	82.38 $\pm$ 0.29	25.56	1.76
MobileNetV3-Small	CNN	69.98 $\pm$ 0.32	71.17 $\pm$ 0.31	70.36 $\pm$ 0.19	70.20 $\pm$ 0.23	2.54	1.08
EfficientNet-B0	CNN	88.88 $\pm$ 0.22	89.11 $\pm$ 0.19	86.53 $\pm$ 0.41	87.96 $\pm$ 0.28	5.29	1.42
ConvNeXt-T	CNN	90.35 $\pm$ 0.28	88.81 $\pm$ 0.29	89.84 $\pm$ 0.23	89.13 $\pm$ 0.34	28.59	2.18
ViT-B-16	Transformer	91.03 $\pm$ 0.22	90.28 $\pm$ 0.34	91.23 $\pm$ 0.28	90.22 $\pm$ 0.27	86.57	3.86
Swin-T	Transformer	90.83 $\pm$ 0.26	88.94 $\pm$ 0.17	90.98 $\pm$ 0.29	90.63 $\pm$ 0.19	28.29	3.24
TransFG	Transformer-FGVC	92.15 $\pm$ 0.35	92.28 $\pm$ 0.23	91.53 $\pm$ 0.20	91.98 $\pm$ 0.18	86.61	3.58
DRS-ViT	Ours	93.93 $\pm$ 0.21	94.68 $\pm$ 0.17	93.29 $\pm$ 0.12	93.91 $\pm$ 0.19	92.4	4.62

Figure 15 presents a representative misclassified sample. The specimen was confirmed by experts as male, but its head was lost during trapping. The SWV response map shows that the model responses remained concentrated on insect morphological regions, including the wing margins and remaining body structures, rather than on the background. Because the head was missing, morphology associated with that region was unavailable, and the specimen was classified as female. The case shows that SWV responses remained localized to insect morphology, while specimen damage and loss of sex-related features coincided with misclassification.

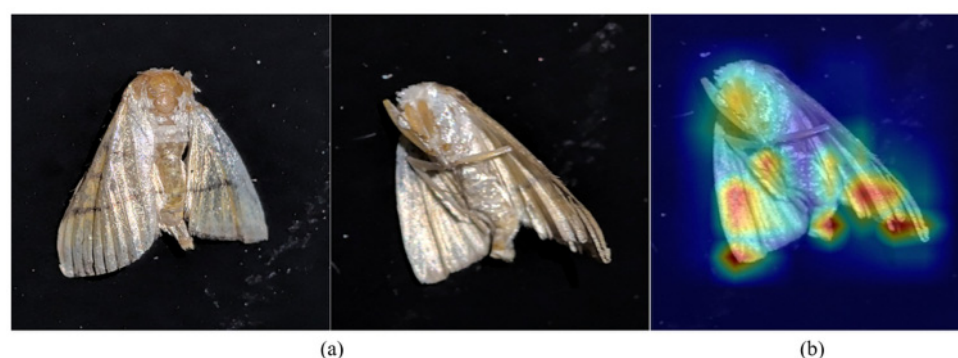


Figure 15. Representative sex-classification failure involving a naturally damaged male *Chilo suppressalis*. (a) Two views of the expert-confirmed male specimen showing the missing head region. (b) Corresponding SWV response map. The specimen was classified as female.

3.2.2. DRS-ViT Ablation Study

As shown in Table 6, the ViT baseline achieved a mean accuracy of $91.03 \pm 0.22\%$. Introducing Convolutional Patch Embedding (CPE), Discriminative-Region Soft-Weighting (SWV), and KANsformer individually resulted in mean accuracy values of $92.02 \pm 0.35\%$, $93.42 \pm 0.26\%$, and $91.73 \pm 0.43\%$, corresponding to gains of 0.99, 2.39, and 0.70 percentage points, respectively. SWV produced the largest individual improvement in mean accuracy among the three modules. Combining CPE and SWV resulted in a mean accuracy of $93.48 \pm 0.28\%$. After integrating all three modules, DRS-ViT achieved mean precision, recall, and accuracy values of $94.68 \pm 0.17\%$, $93.29 \pm 0.12\%$, and $93.93 \pm 0.21\%$, respectively, representing gains of 4.40, 2.06, and 2.90 percentage points over the baseline. The complete configuration achieved the highest mean performance among the evaluated ablation configurations, supporting complementary contributions from the three modules under the tested conditions.

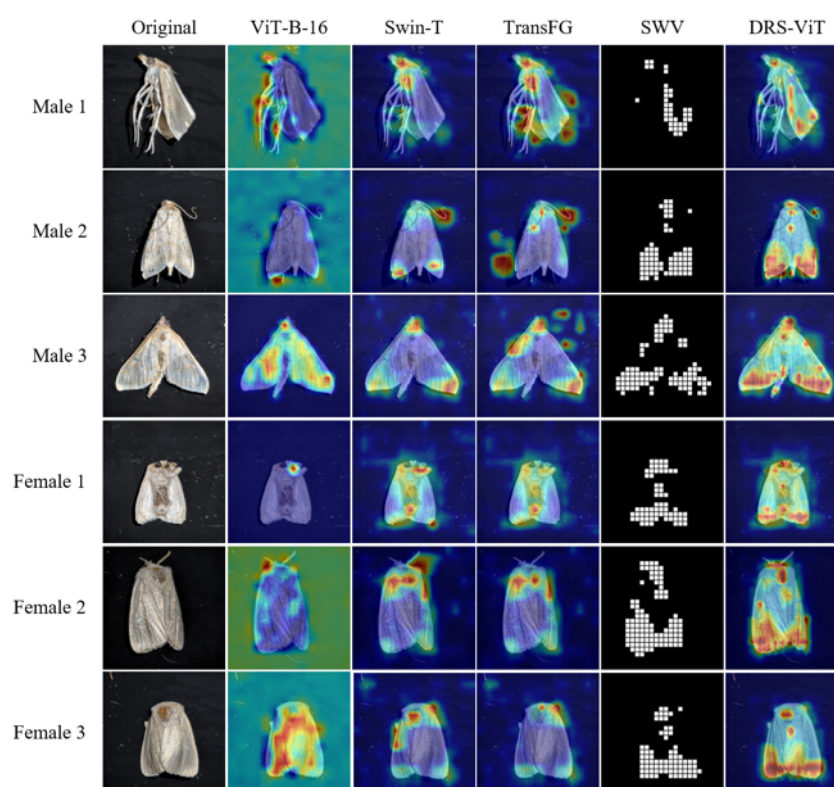
Table 6. Ablation results of DRS-ViT.

NO.	CPE	SWV	Kansformer	P/%	R/%	Acc/%
Baseline	×	×	×	90.28 ± 0.34	91.23 ± 0.28	91.03 ± 0.22
1	✓	×	×	90.74 ± 0.41	88.99 ± 0.27	92.02 ± 0.35
2	×	✓	×	92.48 ± 0.29	91.40 ± 0.44	93.42 ± 0.26
3	×	×	✓	92.16 ± 0.37	91.10 ± 0.25	91.73 ± 0.43
4	✓	✓	×	93.15 ± 0.31	92.49 ± 0.40	93.48 ± 0.28
5	✓	×	✓	92.85 ± 0.46	92.47 ± 0.34	92.89 ± 0.30
6	×	✓	✓	93.72 ± 0.33	92.89 ± 0.23	93.47 ± 0.42
7	✓	✓	✓	94.68 ± 0.17	93.29 ± 0.12	93.93 ± 0.21

“×” indicates that the corresponding module was not included, whereas “✓” indicates that the corresponding module was included.

3.2.3. Attention Visualization

To further explain the response distribution of DRS-ViT, this study visualized and compared the attention responses of different models on male and female samples of the rice stem borer. As shown in Figure 16, the responses of ViT-B-16 were relatively scattered, and some high-response regions deviated from the main insect body and were still affected by background texture interference. Swin-T and TransFG could focus on the insect body region to some extent, but local bias remained in their high-response regions. In contrast, DRS-ViT responses were mainly located around wing margins, antennae, the head, wing-surface textures, and the body axis, with weaker background responses. These maps provide qualitative information about response distribution.

**Figure 16.** Attention-response visualization results.

3.3. Field Data Transmission Performance Analysis

As shown in Table 7, the image acquisition device supported the continuous workflow of pest trapping, image acquisition, edge-side detection, and cloud uploading under actual rice-field conditions. Of the 7168 field-acquired images, 5993 (83.6%) contained at least one

insect and were retained for downstream analysis, whereas the remaining 1175 (16.4%) were empty-tray images acquired during insect-free operating periods and were excluded rather than counted as upload failures. The image upload loss rate was calculated separately as the difference between the number of images sent by the edge terminal and received by the cloud server divided by the number sent. The observed upload loss rate of 8.3% resulted from temporary interruptions in the field 5G network. The upload success rate and latency indicate that continuous edge-to-cloud transmission and remote pest analysis were generally feasible under the tested field conditions.

Table 7. Operation performance of edge–cloud cooperative system in the field experiment.

Metric	Value
Field-collected images	7168
Images containing insects	5993
Insect-containing image rate (%)	83.6
Empty-tray images	1175
Empty-tray image rate (%)	16.4
Average edge-side processing time (ms)	10.6
Average ROI upload latency (s)	0.75
Image upload success rate (%)	91.7
Image upload loss rate (%)	8.3

3.4. End-to-End Classification Performance Analysis

After field operation stability was verified, the end-to-end performance of the edge–cloud collaborative workflow was evaluated using the joint criterion defined in Section 2.7. Among the 7168 images acquired in the field, 5993 contained an expert-confirmed rice stem borer individual and were included in the evaluation; the remaining 1175 were empty-tray images. The evaluated individuals comprised 2397 females and 3596 males. The complete workflow achieved a system-level accuracy of 87.9%, as shown in Table 8.

Table 8. Comparison of end-to-end sex classification performance of different system.

NO.	Edge-Side Model	ROI Foreground Enhancement	Cloud-Based Model	System-Acc/%
1	YOLOv11n	×	ViT-B-16	68.5
2	YOLOv11n	✓	ViT-B-16	75.8
3	YOLOv11n	×	DRS-ViT	76.9
4	YOLOv11n	✓	DRS-ViT	82.4
5	YOLO-CSNet	×	ViT-B-16	73.2
6	YOLO-CSNet	✓	ViT-B-16	80.6
7	YOLO-CSNet	×	DRS-ViT	82.1
8	YOLO-CSNet	✓	DRS-ViT	87.9

“×” indicates that the corresponding module was not included, whereas “✓” indicates that the corresponding module was included.

As shown in Table 8, paired comparisons were conducted while holding the other two processing stages constant. ROI foreground enhancement increased system-level accuracy by 5.5–7.4 percentage points across the evaluated configurations. Replacing ViT-B-16 with DRS-ViT produced gains of 6.6–8.9 percentage points, whereas replacing YOLO11n with YOLO-CSNet produced gains of 4.7–5.5 percentage points. The complete configuration achieved the highest system-level accuracy among the evaluated configurations, reaching 87.9% under the tested field conditions. These gains were not interpreted as additive because errors may propagate and interact across processing stages. Nevertheless, the

paired comparisons indicate that target localization, ROI foreground enhancement, and fine-grained classification make complementary, stage-specific contributions to the final system performance.

To further assess sex-specific classification performance and potential bias, row-normalized confusion matrices were analyzed for DRS-ViT and the complete field workflow, as shown in Figure 17. Under controlled conditions, the recalls for male and female rice stem borers were 95% and 93%, respectively, and the corresponding false-positive rates were 7% and 5%. Under field conditions, the recalls for male and female rice stem borers were 87% and 89%, respectively, with corresponding false-positive rates of 11% and 13%. These results indicate that both DRS-ViT and the complete field workflow maintained relatively balanced classification performance for the two sexes, with no pronounced sex-specific bias under the evaluated conditions.

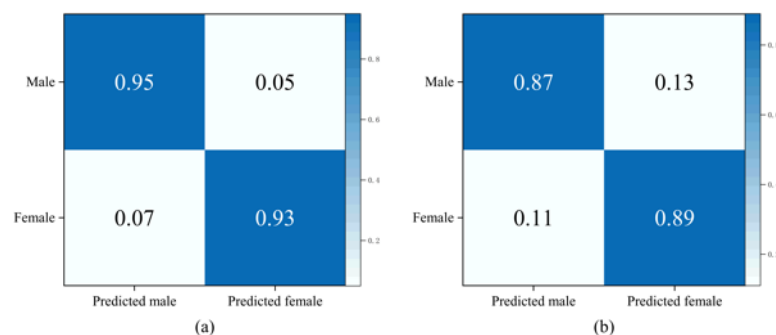


Figure 17. Row-normalized confusion matrices for rice stem borer sex classification. (a) DRS-ViT results on the controlled test set. (b) Final sex-output results of the complete edge-cloud workflow on the temporally independent 2026 field-validation dataset.

4. Discussion

The contributions of this study are methodological and system-level. At the system level, detection, ROI cropping, and foreground enhancement are performed at the edge terminal, whereas fine-grained sex classification is performed in the cloud. Methodologically, the Discriminative-Region Soft-Weighting (SWV) module integrates multi-head voting, continuous attention intensity, spatial smoothing, and residual recalibration to preserve multiple weak sex-related cues rather than only the highest-response regions. This allocation extends previous smart-trap systems [18–20] from automated detection and counting to within-species sex classification under the computational and communication constraints of field devices.

YOLO-CSNet achieved mean mAP@0.5 and mAP@0.5:0.95 values of 95.32% and 73.95%, respectively, on the Rice Insect dataset, and the ablation results indicated complementary contributions from GAM and ASFF, consistent with rice-pest detectors that use attention and multiscale fusion to address small targets, occlusion, and background interference [14–17]. Although these aggregate metrics characterize detection performance under different localization criteria, they do not fully reflect failure behavior under heterogeneous or degraded imaging conditions. Recent multi-domain agricultural detection research has shown that models with comparable nominal accuracy may exhibit different error patterns under spectral variation, occlusion, and image degradation [39]. Unlike open-field imaging, the present system uses an annular shadowless light source, light-absorptive inner surfaces, and a push-pull insect-removal mechanism to reduce illumination variation, reflections, background interference, and insect accumulation before detection. These measures reduce acquisition-induced variation and improve the consistency of images supplied to YOLO-CSNet, although residual image degradation, insect occlusion, and specimen damage may still affect field performance.

ROI foreground enhancement further suppressed tray textures, shadows, and non-target regions, and its gains in the paired end-to-end comparisons indicate that preprocessing contributed to the final system performance. In the cloud-based classification stage, DRS-ViT achieved a mean accuracy of 93.93% and outperformed the evaluated CNN, Transformer, and fine-grained classification baselines under matched conditions. This extends previous pest-recognition research from interspecies classification [21–24] to within-species sex classification. The ablation results support complementary roles for convolutional patch embedding, discriminative-region soft weighting, and KANsformer in local feature extraction, discriminative-region weighting, and global relationship modeling. Attention responses were concentrated around wing margins, the body axis, abdominal textures, and other morphologically relevant regions.

Rice stem borers and rice leafrollers predominated during deployment, but other phototactic insects were occasionally present. YOLO-CSNet functioned as a closed-set species filter, and only ROIs classified as rice stem borers were forwarded for sex classification. Morphologically similar, unmodeled insects could still be assigned to a trained class and enter the downstream workflow as false positives. Their contribution to field errors was not quantified separately. The reported end-to-end accuracy therefore represents performance on expert-confirmed rice stem borer targets under the tested mixed-catch conditions, rather than general rejection of all non-target taxa.

Using a joint criterion that required an IoU of at least 0.7 and correct sex classification, the independent field evaluation achieved a system-level accuracy of 87.9%. All field-validation images were acquired after model development and excluded from training, hyperparameter optimization, threshold selection, and model selection, ensuring temporal independence. However, validation covered only one site and the May–June 2026 deployment period. The 6.03-percentage-point difference between standalone DRS-ViT accuracy (93.93%) and end-to-end system accuracy (87.9%) should not be attributed solely to classifier degradation because the evaluations used different inputs and success criteria. Controlled images contained complete specimens across a broader range of orientations, whereas the field system generally captured one uncontrolled view per individual, which may have left some sex-related features insufficiently visible. System-level accuracy was also affected by localization errors, incomplete ROI cropping, residual background interference, and classification errors. The paired comparisons in Table 8 indicate that errors can propagate and interact across the detection, foreground-processing, and classification stages. Therefore, standalone classifier evaluation may overestimate complete-workflow performance under field conditions.

The field image upload success rate of 91.7% indicates that edge-to-cloud transmission was generally feasible under the tested 5G conditions. Because unsuccessfully uploaded images did not enter the accuracy calculation, upload loss reduced image availability rather than directly affecting the reported system-level accuracy. Nevertheless, the remaining upload failures indicate that communication reliability should be considered in long-term operation. Although the confusion matrices showed no pronounced sex-specific bias under the evaluated conditions, unequal class distributions may cause overall accuracy to obscure asymmetric errors. Because sex-ratio monitoring is an important application of the proposed system [25,26], future multi-site validation should determine whether these class-specific error patterns remain stable and quantify deviations between predicted and manually verified sex ratios. Future work will also extend field validation across sites and seasons, evaluate residual image degradation, occlusion or damage, viewing-angle variation, and device versions, and investigate additional local insect classes, open-set rejection, multi-view acquisition, and multi-pest time-series forecasting.

5. Conclusions

The proposed edge–cloud workflow integrates field image acquisition, edge-side detection and ROI foreground enhancement, and cloud-based sex classification of the rice stem borer. YOLO-CSNet achieved mean mAP@0.5 and mAP@0.5:0.95 values of 95.32% and 73.95%, respectively, while DRS-ViT achieved 93.93% mean accuracy under controlled conditions. The temporally held-out 2026 deployment achieved 87.9% system-level accuracy and a 91.7% image-upload success rate, supporting workflow feasibility at the Shandong site during May–June 2026. Generalizability across sites, seasons, pest densities, environmental backgrounds, and device versions remains to be established. Future work will prioritize broader field validation and extend the system to multi-pest time-series forecasting and population-dynamics modeling.

Author Contributions: Conceptualization, S.Y., L.W. (Luyue Wang) and S.L.; methodology, S.Y. and S.L.; validation, S.Y. and L.W. (Luyue Wang); formal analysis, S.L. and P.M.; investigation, Y.Z. (Yingchao Zhan); data curation, M.L. and Y.Z. (Yige Zheng); writing—original draft preparation, S.Y.; writing—review and editing, S.Y., L.W. (Lixing Wei) and S.L.; visualization, S.Y.; supervision, W.Z.; project administration, P.M.; funding acquisition, S.L. and W.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Shandong Modern Agricultural Industrial Technology System Rice Agricultural Machinery Post Expert Project, grant number SDAIT-17-08.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: The authors gratefully acknowledge Shandong Agricultural University for equipment support and Jinan Wuzhou Urban Agriculture Co., Ltd. for access to the experimental site. The authors also thank Xianliang Hu of Shandong Xiangchen Technology Group Co., Ltd. for providing the baseline pest-monitoring equipment used in this study.

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

References

1. Yuan, S.; Stuart, A.; Laborte, A.; Edreira, J.; Dobermann, A.; Kien, L.; Thúy, L.; Paothong, K.; Traesang, P.; Tint, K.; et al. Southeast Asia must narrow down the yield gap to continue to be a major rice bowl. *Nat. Food* **2022**, *3*, 217–226. [[CrossRef](#)] [[PubMed](#)]
2. Dong, C.; Abil, E.; Ji, R.; Du, Y.; Lu, M. Transcriptomic Analysis of *Chilo suppressalis* (Walker) (Lepidoptera: Pyralidae) Reveals Cold Tolerance Mechanisms Under Parasitism Stress. *Insects* **2025**, *16*, 907. [[CrossRef](#)] [[PubMed](#)]
3. Wang, C.; Wang, X.; Jin, Z.; Müller, C.; Pugh, T.; Chen, A.; Wang, T.; Huang, L.; Zhang, Y.; Li, L.; et al. Occurrence of crop pests and diseases has largely increased in China since 1970. *Nat. Food* **2022**, *3*, 57–65. [[CrossRef](#)] [[PubMed](#)]
4. Ngugi, L.C.; Abellwahab, M.; Abo-Zahhad, M. Recent advances in image processing techniques for automated leaf pest and disease recognition—A review. *Inf. Process. Agric.* **2021**, *8*, 27–51. [[CrossRef](#)]
5. Preti, M.; Verheggen, F.; Angeli, S. Insect pest monitoring with camera-equipped traps: Strengths and limitations. *J. Pest Sci.* **2021**, *94*, 203–217. [[CrossRef](#)]
6. Lima, M.; Leandro, M.; Valero, C.; Coronel, L.; Bazzo, C. Automatic Detection and Monitoring of Insect Pests-A Review. *Agriculture* **2020**, *10*, 161. [[CrossRef](#)]
7. Pecenka, J.; Ingwell, L.; Foster, R.; Krupke, C.; Kaplan, I. IPM reduces insecticide applications by 95% while maintaining or enhancing crop yields through wild pollinator conservation. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2108429118. [[CrossRef](#)] [[PubMed](#)]
8. Wu, Q.; Zeng, J.; Wu, K. Research and application of crop pest monitoring and early warning technology in China. *Eng. Agric.* **2022**, *9*, 19. [[CrossRef](#)]
9. Ariza-Sentís, M.; Vélez, S.; Martínez-Peña, R.; Baja, H.; Valente, J. Object detection and tracking in Precision Farming: A systematic review. *Comput. Electron. Agric.* **2024**, *219*, 108757. [[CrossRef](#)]
10. Tian, H.; Wang, T.; Liu, Y.; Qiao, X.; Li, Y. Computer vision technology in agricultural automation—A review. *Inf. Process. Agric.* **2020**, *7*, 1–19. [[CrossRef](#)]

11. Kasinathan, T.; Singaraju, D.; Uyyala, S.R. Insect classification and detection in field crops using modern machine learning techniques. *Inf. Process. Agric.* **2021**, *8*, 446–457. [[CrossRef](#)]
12. Tian, Y.; Wang, S.; Li, E.; Yang, G.; Liang, Z.; Tan, M. MD-YOLO: Multi-scale Dense YOLO for small target pest detection. *Comput. Electron. Agric.* **2023**, *213*, 108233. [[CrossRef](#)]
13. Zhang, W.; Huang, H.; Sun, Y.; Wu, X. AgriPest-YOLO: A rapid light-trap agricultural pest detection method based on deep learning. *Front. Plant Sci.* **2022**, *13*, 1079384. [[CrossRef](#)] [[PubMed](#)]
14. Yin, J.; Zhu, J.; Chen, G.; Jiang, L.; Zhan, H.; Deng, H.; Long, Y.; Lan, Y.; Wu, B.; Xu, H. An Intelligent Field Monitoring System Based on Enhanced YOLO-RMD Architecture for Real-Time Rice Pest Detection and Management. *Agriculture* **2025**, *15*, 798. [[CrossRef](#)]
15. Qiang, J.; Zhao, L.; Wang, H.; Xu, T.; Jia, Q.; Sun, L. YOLO-PEST: A novel rice pest detection approach based on YOLOv5s. *Plant Methods* **2025**, *21*, 121. [[CrossRef](#)] [[PubMed](#)]
16. Zheng, Y.; Zheng, W.; Du, X. Paddy-YOLO: An accurate method for rice pest detection. *Comput. Electron. Agric.* **2025**, *238*, 110777. [[CrossRef](#)]
17. Yuan, S.; Duan, Y.; Su, H.; Zhou, X.; Hao, Y. HDA-YOLO: A hierarchical and densely-fused attention network for rice pest detection in complex agricultural environments. *Front. Plant Sci.* **2026**, *17*, 1763650. [[CrossRef](#)] [[PubMed](#)]
18. Sun, Y.; Liu, X.; Yuan, M.; Ren, L.; Wang, J.; Chen, Z. Automatic in-trap pest detection using deep learning for pheromone-based *Dendroctonus valens* monitoring. *Biosyst. Eng.* **2018**, *176*, 140–150. [[CrossRef](#)]
19. Hinojosa-Dávalos, J.; Robles-García, M.; Gutiérrez-Lomelí, M.; Jiménez, A.; Lúa, C. Neural Network-Guided Smart Trap for Selective Monitoring of Nocturnal Pest Insects in Agriculture. *Agriculture* **2025**, *15*, 1562. [[CrossRef](#)]
20. Wang, R.-F.; Cui, K.; Schardong, I.B.; Bauer, M.C.; Somala, R.V.; Xu, M.; West, D.; Jones, D.C.; Taylor, S.V.; Roberts, P.M.; et al. JassidNet as a quantization-aware lightweight phenotyping framework for high-throughput cotton jassid (*Amrasca biguttula*) detection and counting toward objective resistance screening. *Ind. Crops Prod.* **2026**, *249*, 123716. [[CrossRef](#)]
21. Fu, X.; Ma, Q.; Yang, F.; Zhang, C.; Zhao, X.; Chang, F.; Han, L. Crop pest image recognition based on the improved ViT method. *Inf. Process. Agric.* **2024**, *11*, 249–259. [[CrossRef](#)]
22. Fang, M.; Tan, Z.; Tang, Y.; Chen, W.; Huang, H.; Dananjayan, S.; He, Y.; Luo, S. Pest-ConFormer: A hybrid CNN-Transformer architecture for large-scale multi-class crop pest recognition. *Expert Syst. Appl.* **2024**, *255*, 124833. [[CrossRef](#)]
23. Kim, G.; Son, C.; Lee, S. ROI-aware multiscale cross-attention vision transformer for pest image identification. *Comput. Electron. Agric.* **2025**, *237*, 110732. [[CrossRef](#)]
24. Zhang, Y.; Chen, L.; Yuan, Y. Multimodal Fine-Grained Transformer Model for Pest Recognition. *Electronics* **2023**, *12*, 2620. [[CrossRef](#)]
25. Chen, R.; Klein, M.; Sheng, C.; Li, Q.; Li, Y.; Li, L.; Hung, X. Mating Disruption or Mass Trapping, Compared with Chemical Insecticides, for Suppression of *Chilo suppressalis* (Lepidoptera: Crambidae) in Northeastern China. *J. Econ. Entomol.* **2014**, *107*, 1828–1838. [[CrossRef](#)] [[PubMed](#)]
26. Cheng, Y.; Hu, M.; Kang, A.; Xiao, Y.; Luo, L.; Jiang, X. The Sex Ratio Indicates the Conclusion and Onset of Population Cycles in the Beet Webworm *Loxostege sticticalis* L. (Lepidoptera: Pyralidae). *Insects* **2023**, *14*, 781. [[CrossRef](#)] [[PubMed](#)]
27. Anami, B.S.; Malvade, N.N.; Palaiah, S. Classification of yield affecting biotic and abiotic paddy crop stresses using field images. *Inf. Process. Agric.* **2020**, *7*, 272–285. [[CrossRef](#)]
28. Sun, H.; Wang, R.-F.; Yu, R.; Sun, Y. Pest image recognition algorithm based on joint adversarial transfer learning. *Measurement* **2026**, *278*, 121620. [[CrossRef](#)]
29. Shimoda, M.; Honda, K. Insect reactions to light and its applications to pest management. *Appl. Entomol. Zool.* **2013**, *48*, 413–421. [[CrossRef](#)]
30. Guo, D.; Zhao, S.Y.; Su, S.C.; Bu, L.A.; Sun, H.; Gao, C.F.; Wu, S.F. Doublesex regulates the development of sex-specific organs in rice stem borer, *Chilo suppressalis*. *Pest Manag. Sci.* **2025**, *81*, 5876–5889. [[CrossRef](#)] [[PubMed](#)]
31. Luo, G.; Yao, J.; Yang, Q.; Zhang, Z.; Hoffmann, A.; Fang, J. Variability in development of the striped rice borer, *Chilo suppressalis* (Lepidoptera: Pyralidae), due to instar number and last instar duration. *Sci. Rep.* **2016**, *6*, 35231. [[CrossRef](#)] [[PubMed](#)]
32. Zahiri, R.; Sarafrazi, A.; Salehi, L.; Kunkel, J.G. A geometric morphometric study on populations of the Rice Stem Borer, *Chilo suppressalis* Walker (Lepidoptera: Crambidae) in northern Iran. *Zool. Middle East* **2006**, *38*, 73–84. [[CrossRef](#)]
33. Ganapathy, M.; Periasamy, S.; Pugazhendi, P.; Antony, C. YOLOv11n for precision agriculture: Lightweight and efficient detection of guava defects across diverse conditions. *J. Sci. Food Agric.* **2025**, *105*, 6182–6195. [[CrossRef](#)] [[PubMed](#)]
34. Liu, Y.; Shao, Z.; Hoffmann, N. Global attention mechanism: Retain information to enhance channel-spatial interactions. *arXiv* **2021**, arXiv:2112.05561. [[CrossRef](#)]
35. Xu, Q.; Wang, J.; Jiang, B.; Luo, B. Fine-Grained Visual Classification via Internal Ensemble Learning Transformer. *IEEE Trans. Multimed.* **2023**, *25*, 9015–9028. [[CrossRef](#)]

36. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; Bengio, Y. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In Proceedings of the 32nd International Conference on Machine Learning, Proceedings of Machine Learning Research, Lille, France, 7–9 July 2015; pp. 2048–2057.
37. Zhu, K.; Wu, J. Residual Attention: A Simple but Effective Method for Multi-Label Recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 184–193.
38. 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*, 1137–1149. [[CrossRef](#)] [[PubMed](#)]
39. Rana, S.; Hensel, O.; Nasirahmadi, A. From vineyard to vision: Multi-domain analysis and mitigation of grape cluster detection failures in complex viticultural environments. *Results Eng.* **2026**, *29*, 108833. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.