IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 16 March 2024, accepted 6 May 2024, date of publication 13 May 2024, date of current version 21 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3400254

RESEARCH ARTICLE

Rice Ears Detection Method Based on Multi-Scale Image Recognition and Attention Mechanism

FEN QIU[1], XIAOJUN SHEN[2,3], CHENG ZHOU[2,3], WUMING HE[2,3], AND LILI YAO[2,3]

[1]Huzhou Academy of Agricultural Sciences, Huzhou 313000, China
[2]School of Information Engineering, Huzhou University, Huzhou 313000, China
[3]Zhejiang Province Key Laboratory of Smart Management and Application of Modern Agricultural Resources, Huzhou 313000, China

Corresponding author: Lili Yao (03120@zjhu.edu.cn)

This work was supported in part by the Grain and Oil Industry Technology Team Project of Zhejiang (Integrated and Demonstrative Project for Hybrid Rice Variety Selection and Biodegradable Mulch Film Matching Film Covering Mechanized Planting Technology), in part by the Natural Science Fund of Huzhou under Grant 2023YZ15, and in part by the Key Research and Development Program of Huzhou under Grant 2022ZD2048.

ABSTRACT Accurate identification of rice ears is crucial for assessing rice yield. Present research mainly relies on single-scale image data for rice ears detection and counting. However, these approaches are susceptible to misdetection and omission due to the intricate environmental conditions in fields. The combination of multi-source images can better overcome the limitations of single-scale images. In this study, based on the YOLOv5s target detection algorithm, a method for rice ears detection and counting applicable to multi-source images is proposed by integrating image data collected by cell phones and UAVs during the rice heading and maturity periods. The proposed method introduces Attention-based Intrascale Feature Interaction (AIFI) to reconstruct the backbone feature extraction network, optimizing feature expression interaction and enhancing handling of the model of advanced semantic information. Additionally, Simplify Optimal Transport Assignment (SimOTA) is employed to achieve a more refined label assignment strategy, thereby optimizing detection performance of the model and addressing difficulties in detecting multiple targets in high-density rice ears environments. Finally, Channel-wise Knowledge Distillation for Dense Prediction (CWD) is utilized to enhance the performance of the model in dense prediction tasks by transferring knowledge between different channels. The experimental results demonstrated good performance of the model on datasets comprising rice at the heading and maturity stages, achieving Precision, Recall, and mAP values of 93%, 85.3%, and 90.3%, respectively. The coefficients of determination (R^2) for the linear fit between test results and the actual statistical results of the model were 0.91, 0.91, 0.90, and 0.88, respectively. The proposed model performs well in the mixed dataset and can be utilized more effectively for accurate identification and counting of rice ears.

INDEX TERMS Rice ears recognition, YOLOv5s, CWD distillation, multi-source images.

I. INTRODUCTION

Rice serves as the primary food source for humankind, and the stabilization of its production holds profound implications for global food security, agricultural economies, and social stability. Ensuring accurate monitoring of rice production is essential for agricultural producers to predict harvests and devise plans for increased yields [1]. The yield of rice and other cereal crops largely depends on the number of ears per unit area. In large-scale production scenarios, traditional manual ears counting methods are not only time-consuming and laborious but also prone to errors [2], [3]. Therefore, there is an urgent need for an efficient and accurate automated ears counting method that meets the requirements of both production and scientific research.

With the rapid development of artificial intelligence and computer vision technology, many researchers have conducted studies on rice ears recognition and counting [4], [5]. Currently, Rice ears detection methods are mainly

The associate editor coordinating the review of this manuscript and approving it for publication was Liandong Zhu.

divided into two categories: image segmentation techniques based on color, texture, and other phenotyping features, and object detection techniques based on deep learning [6]. Reza et al. [7] converted UAV RGB images to LAB color space, and based on color feature differences, rice ears segmentation was performed using K-means clustering and graph cut algorithms. The segmentation relative error ranged from 6% to 33%, and the recognition accuracy was found to be unstable. Hayat et al. [8] proposed an unsupervised Bayesian learning method based on acquired UAV rice ears images, which can classify the rice ears and perform segmentation counting based on the statistical properties of the pixels in the image, achieving an F1 score of 82.1%. Shao et al. [9] combined the LC-FCN [10] model based on transfer learning with the watershed algorithm [11] to achieve field rice ears detection and counting, with an accuracy of 89.88%. The segmentation methods employed in the above-mentioned studies have already achieved basic recognition of rice ears. However, the application of such methods in real field scenarios is relatively limited, with exhibiting weaker robustness and universality. They are mostly applicable only to upright and unobstructed rice ears, making it challenging to identify information from overlapping, irregular shapes, and rice ears in complex backgrounds.

The target detection algorithm, based on deep learning techniques, addresses the limitations of traditional image segmentation methods by utilizing convolutional neural networks (CNN) to extract more advanced feature information (e.g., shape, attitude) of rice ears in paddy field environments [12], [13]. Current deep learning models for target detection are mainly categorized into regression-based single-stage models (e.g., Single Shot MultiBox Detector, You Only Look Once) [14], [15] and candidate region-based two-stage models (e.g., R-CNN, Fast-RCNN) [16], [17]. In the application of the two-stage model, Zhang et al. [18] optimized Fast-RCNN using dilated convolution and proposed a multi-fertility rice ears detection model with a mean Average Precision (mAP) of 80.3%. Xu et al. proposed a multi-scale hybrid window rice ears detection method, MHW-PD, based on convolutional neural networks to enhance the feature richness of rice ears [19]. They combined it with a fusion algorithm to reduce the probability of duplicate model detection frames, achieving an average counting accuracy of 87.2%. Compared to the two-stage model, the single-stage model can directly output position and category information of the target without the need for additional candidate region generation steps, thereby possessing a faster detection speed. Zhou et al. [20] improved the Visual Geometry Group Network (VGGNet) [21] convolutional neural network framework and proposed a region-based fully convolutional network rice ears detection model (R-FCN), which successfully mitigated the model leakage problem by introducing a linear non-maximum suppression method. The model achieved an accuracy of 86.8% on the UAV image test dataset. Sun et al. [22] proposed a curved rice ears detection model based on YOLOv4, utilizing the lightweight MobileNetV2 [23] as the backbone feature extraction network. Convolutional Block Attention Module (CBAM) is introduced in the image feature fusion stage, and soft Non-Maximum Suppression (NMS) is used to address the rice ears occlusion problem, achieving a mAP of 90.32% and a frames per second (FPS) of 44.46. Wang et al. [24] employed a novel method for removing duplicate detection frames to optimize the YOLOv5x model and proposed a rice ears detection model named PanicleDetect. They addressed the issue of missed detection of small target ears by adjusting the resolution of the input image of the model, achieving a mAP of 92.77% and a Mean Absolute Percentage Error (MAPE) of 3.44%. However, most of the above studies only focus on the detection and application of single-scale image data, which is prone to misdetection and omission when applied to the complex field environment. Additionally, the constructed models have high requirements on the data acquisition method and lack the versatility to be tested on platforms of different scales. These problems restrict the deployment of such methods in actual field environments.

Therefore, in this study, an innovative and generalized rice ears detection model was proposed based on YOLOv5s. By thoroughly optimizing the backbone network and the feature learning strategy, to ensure the model achieve good detection results when dealing with images of rice ears from various sensors and different growth periods. The main contributions are summarized as follows:

1) A new rice ears dataset was established, comprising Handphone and UAV images captured at the heading and maturity stages. This dataset aims to enhance the generality and generalization ability of the model, enabling adaptation to platforms of different scales.
2) The Simplify Optimal Transport Assignment (SimOTA) method is introduced to optimize the label assignment of rice ears. This approach enables a more detailed understanding of the positional information of each rice ears in the image, facilitating efficient identification and localization of each rice ear.
3) The Attention-based Intrascale Feature Interaction (AIFI) module is utilized to reconstruct the backbone feature extraction network of YOLOv5s. This improvement strengthens processing of advanced semantic feature of the model
4) By employing the Channel-wise Knowledge Distillation for Dense Prediction (CWD) distillation method, the student model AOD-YOLO is developed to learn deeper features from the teacher model YOLOv5m. This approach effectively enhances the recognition performance of the student model in dense paddy fields, achieving a balance between lightweight design and accuracy.

TABLE 1. Rice ears dataset.

Dataset		Training Set	Valid Set	Test Set	Total images
Heading stage	Handphone	689	104	80	873
	UAV	743	89	107	939
Maturity stage	Handphone	481	51	59	591
	UAV	235	24	23	282
Total		2148	268	269	2685

II. MATERIALS AND METHODS

A. DATA SOURCES

The dataset for this study was obtained from Nanxun District and Deqing County, Huzhou City, Zhejiang Province, China. To enhance the generalization ability of the model, images of rice heading and maturity stages were captured using three devices: the iPhone 14 Pro, Huawei Mate 50, and DJI Mavic 3 UAV. These images were obtained from diverse angles, heights, light intensities, sharpness levels, and within complex surroundings featuring varying degrees of object occlusion. Subsequently, image enhancement operations such as flipping, translation, panning, noise addition were applied, resulting in a total of 2685 images. During the field experiment, the Handphone was positioned directly above the rice ears and into a 30-degree angle of about 400 $\sim$ 600 mm height focusing imaging. For the iPhone 14 Pro and Huawei Mate50, the image resolutions were 4032 $\times$ 3024 and 4160 $\times$ 3120, respectively. There were 873 pictures taken during the heading stage and 591 pictures during the maturity stage. The UAV flew at a height of 5m above the rice canopy and maintained a uniform speed of 3m/s, obtaining images at 1-second intervals. The image resolution was 5280 $\times$ 3986, with a total of 939 images obtained during the heading stage and 282 images obtained at the maturity stage. All images were stored in JPG format, and the data descriptions are shown in Table 1.

In this study, LabelMe (http://labelme.csail.mit.edu/Release3.0/), a generic image annotation tool for target detection tasks, was utilized to annotate the rice ears images. The primary focus was on annotating the living rice ears, while the rest of the image components were uniformly labeled as background. The generated JSON file is then converted into a TXT file containing class names and rice ears coordinates using Python. The total number of input labels for model training is 208,942. The processed dataset is randomly divided into training, validation, and test sets in a ratio of 8:1:1, and the labeled images are shown in Figure 1.

FIGURE 1. Manual annotation of images for datasets.

B. MODEL STRUCTURE AND OPTIMIZATION

1) YOLOv5s NETWORK

YOLO is a contemporary deep learning algorithm extensively employed in target detection. It detects multiple target boxes and categories in an image simultaneously through a single forward pass network. Compared to traditional two-stage detection algorithms, YOLO captures global image information, utilizes Anchor Boxes to accommodate multi-sized targets, and reduces computational complexity. These characteristics provide it with an advantage in real-time video processing and resource-constrained environments.

YOLOv5 is the fifth iteration of the YOLO series. Compared to more advanced versions like YOLOv6 and YOLOv7, YOLOv5 demonstrates significant advantages in inference speed, model weights, and memory usage, striking a balance between real-time processing and accuracy. Additionally, compared to YOLOv8, YOLOv5 boasts mature deployment, wider usage, and superior performance on CPU. The network

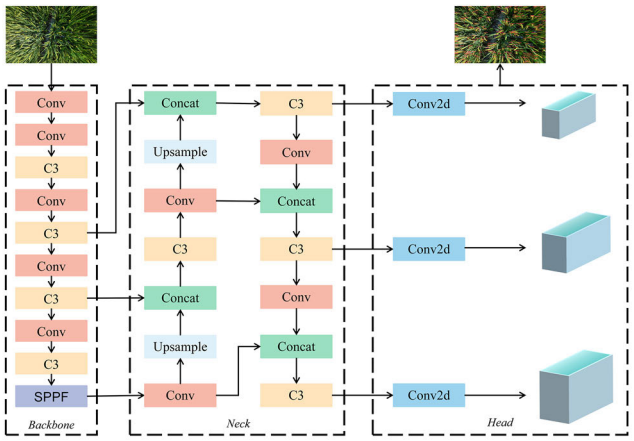

FIGURE 2. The network structure of YOLOv5s.

model structure of YOLOv5 is divided into four versions: YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x.

In this study, YOLOv5s was selected as the benchmark model, suitable for real-time detection tasks on mobile devices due to its fast execution speed and small parameter amount. The network primarily comprises input, backbone feature extraction network, neck network, and head network, with the network structure diagram illustrated in Figure 2. Additionally, three methods, namely SimOTA, AIFI, and CWD distillation, were employed in this study to optimize the YOLOv5s network.

2) SIMPLIFY OPTIMAL TRANSPORT ASSIGNMENT

In the process of rice ears recognition using the YOLOv5s model, the utilization of predefined anchor frames poses challenges in accurately assigning relevant prediction frames to true labels in complex surroundings. This issue becomes more noticeable for rice ears exhibiting varying sizes, postures, and levels of occlusion. The traditional static label assignment strategy with fixed thresholds often results in two or more positive samples in the generated prediction box, which fails to meet the stringent positional accuracy requirements, leading to false detections, omissions, and instances of multiple detections.

To address these issues, the SimOTA method [25] is introduced to optimize the label assignment of rice ears, enabling a more detailed understanding of the positional information of each rice ears in the image. SimOTA approaches the label assignment problem from a global perspective, considering the real box as the supplier and the model-predicted boxes in the training data as the demanders. By defining the unit transportation cost between each demander and supplier, it seeks the optimal allocation scheme for positive and negative samples to minimize the global transportation cost. This approach provides a more flexible and efficient label allocation, thereby improving performance of the model in complex scenarios. Specifically, Anchors are initially utilized for screening, taking into account factors such as location Intersection over Union (IoU), background target, and category. Next, the loss function is calculated to construct cost and IoU matrices, and k candidate boxes are dynamically assigned to each target box based on the top 10 IoU values in the IoU matrix. Finally, the top k candidate boxes are filtered according to the cost matrix, excluding duplicate candidate boxes to identify the best candidate box with the smallest cost value. The flowchart of method is shown Figure 3.

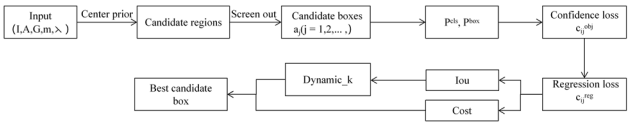

FIGURE 3. SimOTA workflow.

where I represents the input rice ears image, A represents a set of anchor points, G represents the real value labeling of the rice ears in image I, m represents the number of iterations, and λ represents the weighting coefficient with a value of 5. Since the study focuses on single-category target detection, the transportation cost is primarily weighted by the weighted sum of the confidence loss, c_{ij}^{obj}, and the bounding box regression loss, λc_{ij}^{reg}. The calculation method for the cost is shown as Equation (1).

$$c_{ij} = c_{ij}^{obj} + \lambda c_{ij}^{reg} \tag{1}$$

cij represents the transportation cost of an instance of rice ears in the image from the ith real box to the jth prediction box.

3) ATTENTION-BASED INTRASCALE FEATURE INTERACTION

Due to environmental factors and growth conditions, rice ears may fall, droop, and split, resulting in the mechanism of fusing different scale feature information in the SPPF layer of the YOLOv5s backbone network inapplicable.

Therefore, in this study, the AIFI [26] module is introduced to reconstruct the backbone feature extraction network of YOLOv5s. This allows the model to focus more intensively on learning the high-level features of the rice ears (e.g., shape and pose), while reducing the processing of low-level features. As a result, the model can better distinguish between rice ears with different shapes and levels of occlusion. The introduction of this module effectively reduces the power consumption the computational resources of the model, improves robustness of the model in practical applications, and enhances its suitability for the task of rice ears detection in complex environments.

Furthermore, to alleviate the impact of the AIFI module on the weight of the model, this study reconstructs the backbone network of YOLOv5s. A Convolutional layer is introduced before the AIFI layer to reduce the dimension of the feature map, consequently reducing the number of parameters and computational workload of the model. The reconstructed backbone network is shown Figure 4.

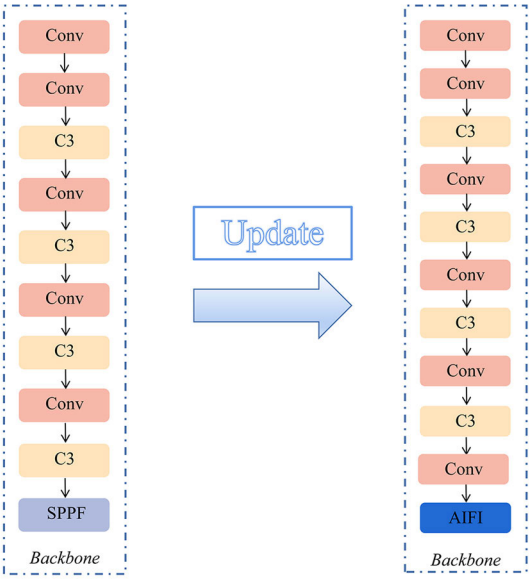

FIGURE 4. Backbone network reconstruction figure.

4) CHANNEL-WISE KNOWLEDGE DISTILLATION FOR DENSE PREDICTION

The rice ears recognition task encounters complex environmental changes, uncertainty in lighting conditions, diversity of rice ears poses, and occlusion. These factors contribute to the poor performance of existing models in addressing this task and make it challenging to achieve accurate rice ears detection. To address the above-mentioned challenges, this study introduces the CWD distillation method [27]. The core idea of CWD distillation involves comparing the channel attention distributions of the teacher network and the student network. These distributions are treated as probability distributions, and the difference between them is measured in terms of KL dispersion, which captures channel information. This allows the student network to better leverage information from the teacher network to enhance its perception and comprehension of fine-grained features. In addition, the method retains high-level semantic information while focusing on detailed features at the channel level. This enhancement improves the accurate recognition of fine structures by the student network, thereby better addressing the problem of dense target recognition in complex scenarios. The loss function calculation method for CWD is shown in Equation (2).

$$L_{CWD} = \frac{T^2}{C} \sum_{c=1}^{C} \sum_{h=1}^{H} \sum_{w=1}^{W} P_{teacher}(h, w, c) \times \log\left(\frac{P_{teacher}(h, w, c)}{P_{student}(h, w, c)}\right) \quad (2)$$

LCWD represents the loss function of CWD, T represents the hyperparameter (temperature), H and W represent the height and width of the feature map, h and w represent the spatial location on the feature map, and C represents the channel index. Pteacher (h, w, c) represents the probability of the network of the teacher having channel c at location (h, w), and Pstudent(h, w, c) represents the probability of the network of the student having channel c at location (h, w).

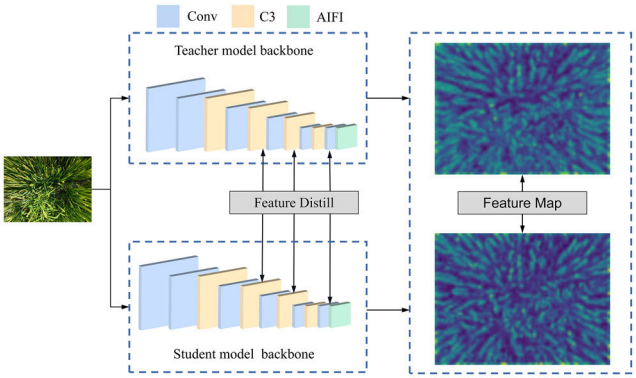

FIGURE 5. CWD distillation flow.

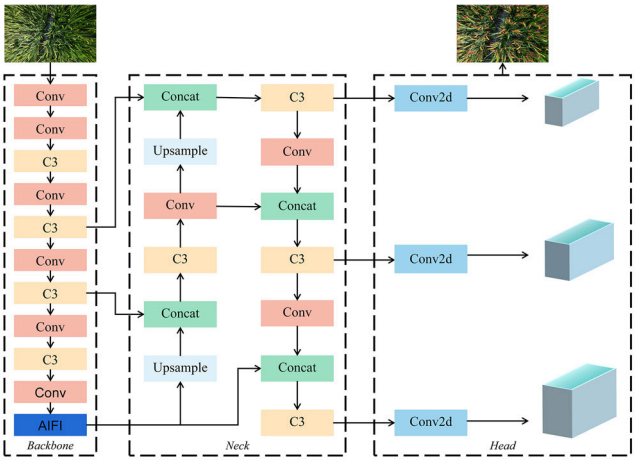

FIGURE 6. AOD-YOLO network structure.

In this study, YOLOv5m is chosen as the teacher network, while the enhanced YOLOv5s network serves as the student network. AIFI is introduced into the teacher network YOLOv5m to align the structure of the backbone feature extraction network between the two models. The distillation process is illustrated in Figure 5.

5) MODEL OPTIMIZATION

Based on the methods of SimOTA, AIFI, and CWD described above, this study proposes an improved rice ears detection model named AOD-YOLO. AOD-YOLO effectively addresses the recognition challenges brought by dense occlusion and shape variations of rice ears in practical paddy field environments.

The model structure is shown in Figure 6. Among them, the AIFI module replaces the SPPF layer in the YOLOv5s backbone network, CSPdarknet53, enhancing the ability of the model to extract high-level semantic features. This enables the model to more accurately capture abstract information related to the rice ears in the image, improving its ability to distinguish rice ears with different shapes and degrees of

occlusion. Simultaneously, the SimOTA method is employed to optimize the label matching strategy of the model, improving its accuracy in label matching. This enables the model to comprehend the location information of each rice ear in the image more intricately, facilitating precise localization and identification of rice ears within the image. Moreover, the CWD method is applied to distill the three output layers of the backbone feature extraction network. This facilitates the transfer of knowledge from the teacher model to the student model. The purpose is to guide the student model in acquiring a more profound understanding of the detailed features of rice ears, thereby enhancing the discriminative ability of the model. This method helps address the challenge of dense rice ears recognition in complex scenes.

C. MODEL PERFORMANCE EVALUATION

The commonly used model evaluation metrics in target detection tasks are mAP, Precision (P), Recall (R), Coefficient of Determination (R^2), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE).

P and R are defined as the ratios of TP to FP and TP to FN, respectively. The IoU value was utilized to determine whether the detection frame matched the rice ears bounding box. Matched detection frames were labeled as true positives (TP), unmatched ones were labeled as false positives (FP), background regions mistakenly detected as rice ears were labeled as true negatives (TN), and undetected rice ears were labeled as false negatives (FN). Average Precision (AP) is a metric that measures the accuracy of an object detection algorithm at different confidence levels, while mAP averages the AP values across all categories. Higher mAP values indicate higher accuracy of the network in the object detection task, serving as a comprehensive indicator of the overall performance of the model. The calculation formulas are as shown in equations (3) to (6).

$$P = \frac{TP}{FP + TP} \tag{3}$$

$$R = \frac{TP}{FN + TP} \tag{4}$$

$$AP = \int_0^1 P(R)\, d(R) \tag{5}$$

$$mAP = \frac{\sum_{k=1}^{n} (AP)_k}{n} \tag{6}$$

Model testing results were validated using the R^2, MAE, and RMSE. The R^2 is used to measure how well the model fits the actual data, the MAE measures the average prediction error, and the RMSE focuses on outliers and overall accuracy. The calculation formulas are as shown in equations (7) to (9).

$$R^2 = 1 - \frac{\sum_{j=1}^{m} \left(S_j - \hat{S}_j\right)^2}{\sum_{j=1}^{m} \sum_j \left(S_j - \bar{S}_j\right)^2} \tag{7}$$

$$MAE = \frac{1}{m} \sum_{j=1}^{m} \left|S_j - \hat{S}_j\right| \tag{8}$$

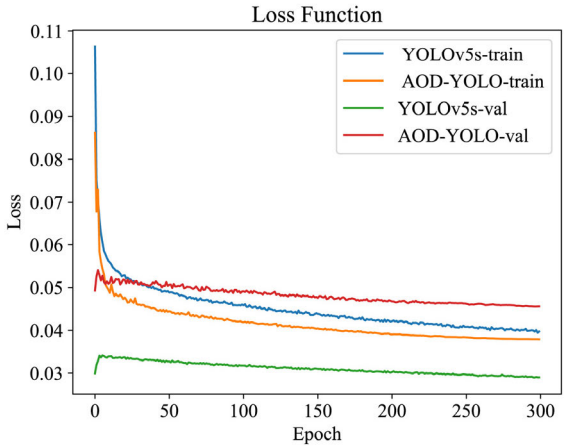

FIGURE 7. The graph of loss function change.

$$RMSE = \sqrt{\frac{\sum_{j=1}^{m} \left(S_j - \hat{S}_j\right)^2}{m}} \tag{9}$$

III. RESULTS

A. EXPERIMENTAL ENVIRONMENT

The hardware platform for training the rice ears recognition model consisted of an AMD EPYC 7742 64-Core Processor with 80 GB RAM and an NVIDIA GeForce RTX 3090 GPU with 24 GB of video memory. The software platform used included PyTorch 1.12.0, CUDA 11.3, and cuDNN 8.2, along with Python 3.8.

The most suitable parameters are determined through fine-tuning during the model training iterations. Stochastic Gradient Descent (SGD) is selected for model parameter optimization. The initial learning rate is set to 0.01, the weight decay value is 0.937, the NMS threshold is 0.45, the confidence threshold is 0.25, and the IoU threshold is 0.5. The cosine annealing method with learning rate decay is employed for 300 training epochs (iteration rounds). The batch size, representing the number of training images in each batch, is set to 32.

As shown in Figure 7, the selected parameters are deemed suitable for model training by observing the figure of the loss function change. With the increase of epochs, the loss curve gradually decreases and stabilizes, indicating that the model has converged.

B. ABLATON EXPERIMENTS

To evaluate the performance of the AOD-YOLO algorithm in rice ears detection, YOLOv5s is utilized as the base model. The improved method is gradually introduced, and the effectiveness of the improved method is verified through ablation experiments. The results are presented in Table 2.

The experimental design is divided into three steps. Firstly, the model label matching strategy is optimized by introducing the SimOTA method. Without changing the model size and computation, the experimental results show

TABLE 2. Ablation experiments.

SimOTA	AIFI	CWD	Weights (MB)	FLOPS (* 10^9)	P (%)	R (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)
—	—	—	13.7	15.8	92.3	82.9	88.4	0.553
√	—	—	13.7	15.8	92.4	83.9	88.9	0.552
√	√	—	14	15.7	92.5	83.9	89.2	0.554
√	√	√	14	15.7	93	85.3	90.3	0.577

TABLE 3. Comparison results of different models based on the same dataset.

Model	Parameters (* 10^6)	Weight (MB)	FLOPs (* 10^9)	P (%)	R (%)	mAP@0.5 (%)
YOLOv7-tiny	6	11.7	13	90	77.3	83.4
YOLOv7	36.4	71.3	103.2	94.8	87.4	90.5
YOLOv8s	11.1	21.4	28.4	86.2	77	84.1
YOLOv8n	3	5.97	8.1	87	77	83.9
AOD-YOLO	7.1	14	15.7	93	85.3	90.3

that the P of the model improves by 0.1%, R by 1%, and mAP by 0.5%, reaching 92.3%, 82.9%, and 88.45%, respectively. Secondly, the AIFI module replaces the SPPF module in the backbone network, and a new convolutional layer is introduced to reconfigure the backbone network. Although the model weights increase slightly by 0.3 MB to 14 MB, the computational effort is reduced by 0.1×10^9 to 15.7×10^9. Accordingly, P and mAP are improved by 0.1% and 0.3% to 92.5% and 89.2%, respectively. Finally, YOLOv5m was used as the teacher model, and the YOLOv5s network with SimOTA and AIFI was employed as the student model, which underwent feature distillation using the CWD method. The experimental results show that P, R, and mAP are improved by 0.5%, 1.4%, and 1.1% to 93%, 85.3%, and 90.3%, respectively. This study also tested the model using the P-R plot curve method for the original network YOLOv5s and the progressively improved network, as shown in Figure 8. The larger the area under the curve (AUC), the better the detection performance of the model. In conclusion, these three optimization steps significantly improve the model performance and effectively reduce the probability of false and missed detections.

C. COMPRISON OF DIFFERENT DETECTION MODELS

Under the premise of ensuring the consistency of the dataset, this study fully demonstrates the good performance of AOD-YOLO by comprehensively comparing it with state-of-the-art models in the YOLO series of current mainstream target detection models. The specific comparison results are shown in Table 3.

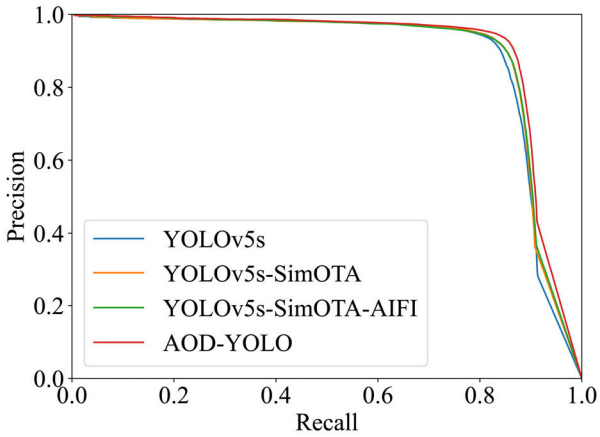

FIGURE 8. Precision - recall diagram.

Compared to YOLOv7 and YOLOv8s, the AOD-YOLO model achieves significant advantages in terms of the number of parameters, computation, and weights. The number of parameters of AOD-YOLO is 7.1×10^6, the computation is 15.7×10^9, and the weights are 14 MB, which are only 19.2% and 63.9% of those of YOLOv7 and YOLOv8s, respectively.

Additionally, the mAP of the AOD-YOLO model is 6.2% higher than that of the YOLOv8s, demonstrating superior detection accuracy at a relatively small model size. Although the mAP of the YOLOv7 model is slightly higher than that of the AOD-YOLO model by 0.2%, the weights, parameter counts, and computation of YOLOv7 model are 5.09, 5.13, and 6.57 times higher than those of the AOD-YOLO, respectively. Compared to YOLOv7-tiny and YOLOv8n, AOD-YOLO demonstrates far superior detection accuracy, with a P of 93%, R of 85.3%, and mAP of 90.3%. These metrics are 0.3%, 8%, and 6.9% higher in AOD-YOLO than in YOLOv7-tiny and YOLOv8n, respectively. Although AOD-YOLO may be slightly inferior in lightweight metrics, its high level of detection accuracy makes it ideal for real-world applications and helps avoid problems of missed and false detections that can occur when models are deployed on mobile devices.

The AOD-YOLO model is compared with the YOLOv7-tiny, YOLOv8n, and YOLOv8s models for real picture detection, using a UAV photo and two Handphone pictures taken from different angles for validation experiments. As shown in Figure 9, the blue detection boxes represent missed detections of rice ears for the YOLOv7-tiny, YOLOv8n, and YOLOv8s models relative to the AOD-YOLO model, while the purple detection boxes represent Mis-checked and over-checked detections of rice ears by these models. Observing the experimental results, it is evident that the AOD-YOLO model exhibits significantly higher detection accuracy in complex field environments. This suggests that the AOD-YOLO model has stronger performance and robustness relative to the YOLOv7-tiny, YOLOv8n, and YOLOv8s models when dealing with the task of rice ears detection under challenging background conditions.

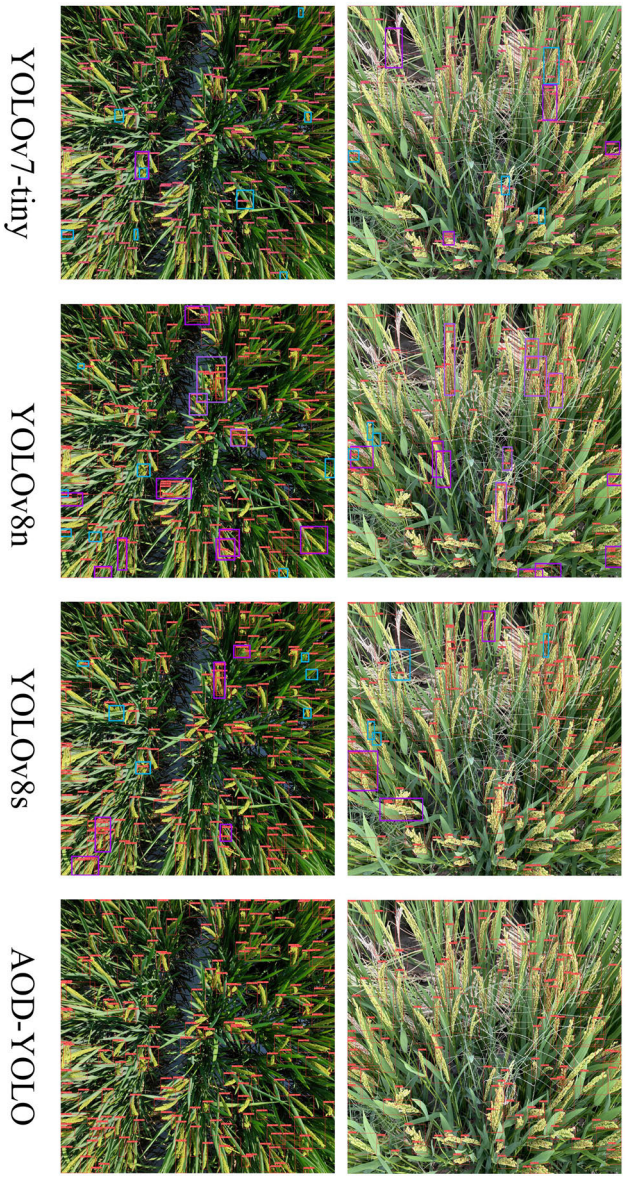

FIGURE 9. A comparison figure of multi-model detection counts. Blue detection boxes represent missed detections of rice ears relative to the AOD-YOLO model. The purple detection boxes represent Mis-checked and over-checked detections of rice ears.

D. MODEL COUNTING VERIFICATION ANALYSIS

To comprehensively verify the robustness, effectiveness, and generalization ability of the trained models in real-world scenarios, this study selects thirty Handphone images including rice heading and maturity stages, as along with UAV images. Each of these images is drawn from the test dataset and serves as an evaluation sample. These field rice ears images will be detected and counted using two models, AOD-YOLO and YOLOv5s, respectively. Linear regression analysis will be used to evaluate the prediction effect of the models and further measure the performance of the models in the detection of rice ears at the heading and maturity stages.

As shown in Table 4 and Figure 10, the R^2 values of the AOD-YOLO model for Handphone and UAV images at the heading and maturity stages are 0.91, 0.90, 0.90, and 0.88, respectively. The MAE values are 3.87, 5.27, 4.73, and 7.29, and the RMSE values are 5.04, 6.95, 6.05, and 9.38, respectively. For the YOLOv5s model, the R^2 values are 0.90, 0.88, 0.88, and 0.87, the MAE values are 4.51, 7.38, 5.96, and 8.32, and the RMSE values are 5.84, 9.12, 7.48, and 10.35, respectively. It can be observed that the prediction performance of the AOD-YOLO model proposed in this study outperforms that of the original YOLOv5s model for all datasets.

Among them, AOD-YOLO has better prediction performance for Handphone and UAV images during the rice heading growth period. This is attributed to the enhanced ability of the model to accurately capture high-level features, facilitated by the regular shape of rice ears and reduced occlusion during the heading period. This aids in delicately identifying the edges and shapes of each rice ears. The prediction performance of AOD-YOLO in Handphone and UAV images at the maturity stage is slightly weaker than that at the heading stage. This is attributed to the fact that the rice

TABLE 4. Accuracy evaluation of test results of different models.

MODEL	STAGE	INSTRUMENT	R^2	MAE	RMSE	MODEL	STAGE	INSTRUMENT	R^2	MAE	RMSE
AOD-YOLO	Heading stage	Handphone	0.91	3.87	5.06	YOLOv5s	Heading stage	Handphone	0.90	4.51	5.84
		UAV	0.90	5.27	6.95			UAV	0.88	7.38	9.12
	Maturity stage	Handphone	0.90	4.73	6.05		Maturity stage	Handphone	0.88	5.96	7.48
		UAV	0.88	7.29	9.38			UAV	0.87	8.32	10.35

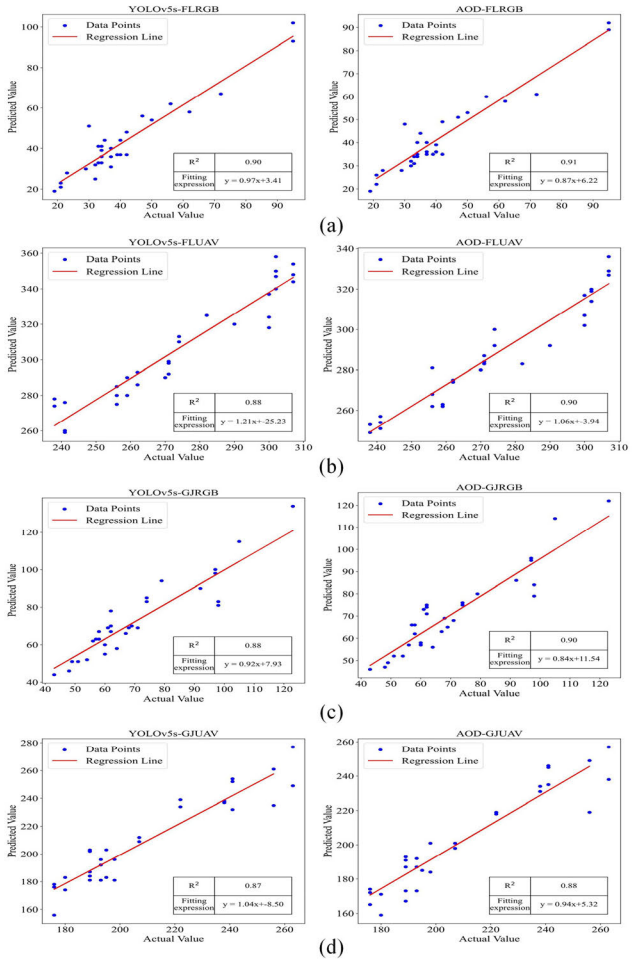

FIGURE 10. The linear fitting results of AOD-YOLO and YOLOv5s for rice ears dataset from different platforms and growth stages are as follows: (a) for the heading stage Handphone dataset, (b) for the heading stage UAV dataset, (c) for the maturity stage Handphone dataset, and (d) for the maturity stage UAV dataset.

ears are dense and the occlusion is increased, resulting in blurred target edges, which makes it difficult for the model to accurately identify and locate each rice ear. Additionally, during this growth period, the overall structure of the rice ears changes because it has been tilted, bent, and split, making it difficult for the model to capture the complete shape of the target. This leads to mis-checked and over-checked, and affects the detection accuracy of the model.

IV. DISCUSSION

Existing studies typically concentrate on processing either UAV or smartphone images but overlook the integration of both, thereby limiting the generalizability of the model. To address this problem, this study proposes a comprehensive rice ears detection method, AOD-YOLO, which aims to enable the model to accurately recognize rice ears in images obtained by platforms with different scales across different rice growth stages.

In addition to the YOLO series of the models, this study also compares the performance of the proposed method with other existing improved rice ears detection methods. Including the improved YOLOv4 detection model for identifying curved rice ears using UAV images proposed by Sun et al. [22], which has a mAP of 89% and an FPS of 44.46. The rice ears counting algorithm RFF-PC proposed by Chen et al. [28], which employs a multiscale convolution (MSConv) and feature pyramid fusion (FPF) strategy and has an average counting accuracy of 89.80%. Zhang et al. [29] proposed an improved rice multiple fertility detection model based on Fast-RCNN, which utilizes Inception_ResNet-v2 to reconstruct the feature extraction network and adopts the feature pyramid network (FPN) and regional proposal network (RPN) fusion to integrate multi-scale semantic information, achieving a mAP of 92.47% and FPS of 4.69. The comprehensive comparison results are shown in Table 5.

According to the comparison results, AOD-YOLO outperforms YOLOv4, which introduces the CBAM attention mechanism and MobileNetV2 lightweight feature extraction network, with a 1.3% improvement in detection accuracy, a 65.54 improvement in FPS, and a detection speed that is almost 2.5 times that of improved-YOLOv4. Relative to RFF-PC, AOD-YOLO improves its detection accuracy by 0.5%. Compared to the improved Faster R-CNN, AOD-YOLO is nearly 27 times ahead in terms of detection speed, but its detection accuracy is slightly reduced by 2.17%. This is due to the fact that improved Faster R-CNN is only applicable to UAV images, which is a relatively simple task.

In summary, AOD-YOLO, a generalized rice ears detection model constructed in this study, maintains satisfactory accuracy and efficient detection speed under bimodal demands. The model is suitable for mobile hardware device deployment and can accomplish real-time detection tasks in complex field environments. The following study will focus on expanding

TABLE 5. Comparing with the methods proposed in other rice ears recognition studies.

Model	mAP@0.5（%）	FPS
improved-YOLOv4	89.0	44.46
RFF-PC	89.8	NA
improved Faster R-CNN	92.47	4.69
AOD-YOLO	90.3	110

the size and diversity of the dataset to improve the detection accuracy and generalization ability of the AOD-YOLO model in complex environments. For the misdetection and omission problems of the generalized model, it is planned to introduce multi-task learning to share the underlying features, cross-modal learning to mitigate the inter-instrument variations, as well as weakly-supervised learning and active learning strategies to improve the accuracy and generalization of the model.

V. CONCLUSION

This study proposes a universal rice ears detection model named AOD-YOLO. The model integrates the AIFI module, SimOTA method, and CWD distillation technology to optimize the backbone network and training strategy of YOLOv5s. This approach addresses the challenge of difficult detection of rice ears under complex environmental conditions and multi-scale scenarios, caused by dense object distribution, significant variations in object sizes, and severe occlusions. AOD-YOLO achieved a mAP of 90.3% in the experiment for rice ears recognition. In the datasets of both handphones and drones during the heading and maturity stages, the R^2 fit between the predicted and actual values of the AOD-YOLO model was 0.91, 0.90, 0.90, and 0.88 respectively. This study conducted comparative tests between AOD-YOLO and various advanced YOLO series detection models, demonstrating the superior detection accuracy of AOD-YOLO. Moreover, this model possesses sufficient versatility and can be easily deployed across different platforms to accomplish real-time rice ears recognition tasks in actual agricultural field environments.

REFERENCES

[1] Q. Zhang, "Strategies for developing green super rice," *Proc. Nat. Acad. Sci. USA*, vol. 104, no. 42, pp. 16402–16409, Oct. 2007.
[2] A. Ferrante, J. Cartelle, R. Savin, and G. A. Slafer, "Yield determination, interplay between major components and yield stability in a traditional and a contemporary wheat across a wide range of environments," *Field Crops Res.*, vol. 203, pp. 114–127, Mar. 2017.
[3] G. A. Slafer, R. Savin, and V. O. Sadras, "Coarse and fine regulation of wheat yield components in response to genotype and environment," *Field Crops Res.*, vol. 157, pp. 71–83, Feb. 2014.
[4] J. Ma, Y. Li, K. Du, F. Zheng, L. Zhang, Z. Gong, and W. Jiao, "Segmenting ears of winter wheat at flowering stage using digital images and deep learning," *Comput. Electron. Agricult.*, vol. 168, Jan. 2020, Art. no. 105159.
[5] X. Xiong, L. Duan, L. Liu, H. Tu, P. Yang, D. Wu, G. Chen, L. Xiong, W. Yang, and Q. Liu, "Panicle-SEG: A robust image segmentation method for rice panicles in the field based on deep learning and superpixel optimization," *Plant Methods*, vol. 13, no. 1, pp. 1–15, Dec. 2017.
[6] Y. Yuan, L. Chen, H. Wu, and L. Li, "Advanced agricultural disease image recognition technologies: A review," *Inf. Process. Agricult.*, vol. 9, no. 1, pp. 48–59, Mar. 2022.
[7] M. N. Reza, I. S. Na, S. W. Baek, and K.-H. Lee, "Rice yield estimation based on K-means clustering with graph-cut segmentation using low-altitude UAV images," *Biosyst. Eng.*, vol. 177, pp. 109–121, Jan. 2019.
[8] M. A. Hayat, J. Wu, and Y. Cao, "Unsupervised Bayesian learning for rice panicle segmentation with UAV images," *Plant Methods*, vol. 16, no. 1, pp. 1–13, Dec. 2020.
[9] H. Shao, R. Tang, Y. Lei, J. Mu, Y. Guan, and Y. Xiang, "Rice ear counting based on image segmentation and establishment of a dataset," *Plants*, vol. 10, no. 8, p. 1625, Aug. 2021.
[10] I. H. Laradji, N. Rostamzadeh, P. O. Pinheiro, D. Vazquez, and M. Schmidt, "Where are the blobs: Counting by localization with point supervision," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 547–562.
[11] A. Kornilov and I. Safonov, "An overview of watershed algorithm implementations in open source libraries," *J. Imag.*, vol. 4, no. 10, p. 123, Oct. 2018.
[12] M. Zhang, H. Lin, G. Wang, H. Sun, and J. Fu, "Mapping paddy rice using a convolutional neural network (CNN) with Landsat 8 datasets in the Dongting lake area, China," *Remote Sens.*, vol. 10, no. 11, p. 1840, Nov. 2018.
[13] A. Kamilaris and F. X. Prenafeta-Boldú, "Deep learning in agriculture: A survey," *Comput. Electron. Agricult.*, vol. 147, pp. 70–90, Apr. 2018.
[14] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, "A review of YOLO algorithm developments," *Proc. Comput. Sci.*, vol. 199, pp. 1066–1073, Jan. 2022.
[15] C. Ning, H. Zhou, Y. Song, and J. Tang, "Inception single shot MultiBox detector for object detection," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2017, pp. 549–554.
[16] J. Xu, H. Ren, S. Cai, and X. Zhang, "An improved faster R-CNN algorithm for assisted detection of lung nodules," *Comput. Biol. Med.*, vol. 153, Feb. 2023, Art. no. 106470.
[17] P. Bharati and A. Pramanik, "Deep learning techniques-R-CNN to mask R-CNN: A survey," in *Proc. CIPR*, 2019, pp. 657–668.
[18] Y. Zhang, D. Xiao, H. Chen, and Y. Liu, "Rice panicle detection method based on improved faster R-CNN," *Trans. Chin. Soc. Agricult. Mach.*, vol. 52, pp. 231–240, Jan. 2021.
[19] C. Xu, H. Jiang, P. Yuen, K. Zaki Ahmad, and Y. Chen, "MHW-PD: A robust rice panicles counting algorithm based on deep learning and multi-scale hybrid window," *Comput. Electron. Agricult.*, vol. 173, Jun. 2020, Art. no. 105375.
[20] C. Zhou, H. Ye, J. Hu, X. Shi, S. Hua, J. Yue, Z. Xu, and G. Yang, "Automated counting of rice panicle by applying deep learning model to images from unmanned aerial vehicle platform," *Sensors*, vol. 19, no. 14, p. 3106, Jul. 2019.
[21] A. Vedaldi and A. Zisserman, "VGG convolutional neural networks practical," Dept. Eng. Sci., Univ. Oxford, 2016, vol. 66.
[22] B. Sun, W. Zhou, S. Zhu, S. Huang, X. Yu, Z. Wu, X. Lei, D. Yin, H. Xia, Y. Chen, F. Deng, Y. Tao, H. Cheng, X. Jin, and W. Ren, "Universal detection of curved rice panicles in complex environments using aerial images and improved YOLOv4 model," *Frontiers Plant Sci.*, vol. 13, Nov. 2022, Art. no. 1021398.
[23] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.
[24] X. Wang, W. Yang, Q. Lv, C. Huang, X. Liang, G. Chen, L. Xiong, and L. Duan, "Field rice panicle detection and counting based on deep learning," *Frontiers Plant Sci.*, vol. 13, Aug. 2022, Art. no. 966495.
[25] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," 2021, *arXiv:2107.08430*.
[26] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, "DETRs beat YOLOs on real-time object detection," 2023, *arXiv:2304.08069*.
[27] C. Shu, Y. Liu, J. Gao, Z. Yan, and C. Shen, "Channel-wise knowledge distillation for dense prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 5291–5300.

[28] Y. Chen, R. Xin, H. Jiang, Y. Liu, X. Zhang, and J. Yu, "Refined feature fusion for in-field high-density and multi-scale rice panicle counting in UAV images," *Comput. Electron. Agricult.*, vol. 211, Aug. 2023, Art. no. 108032.
[29] Y. Zhang, D. Xiao, Y. Liu, and H. Wu, "An algorithm for automatic identification of multiple developmental stages of rice spikes based on improved faster R-CNN," *Crop J.*, vol. 10, no. 5, pp. 1323–1333, Oct. 2022.

FEN QIU received the bachelor's degree from Zhejiang A&F University. Currently, she is with Huzhou Academy of Agricultural Sciences. Her research interests include plastic film mulching cultivation of rice, acquisition of rice growth information, and monitoring of rice quality.

XIAOJUN SHEN received the bachelor's degree from Guizhou Institute of Technology. He is currently pursuing the degree with the School of Information Engineering, Huzhou University, with a focus on smart agriculture and agricultural machine vision.

CHENG ZHOU received the Ph.D. degree from Northeast Agricultural University. He is currently with the School of Information Engineering, Huzhou University, with a focus on variable rate fertilization and precision seeding equipment.

WUMING HE received the master's degree from Guangxi Normal University. Currently, he is with the School of Information Engineering, Huzhou University. His research interests include intelligent information processing and embedded systems.

LILI YAO received the Ph.D. degree from Nanjing Agricultural University. He is currently with the School of Information Engineering, Huzhou University. His research interests include the Internet of Things, field sensors, and agricultural machine vision.