European Journal of Agronomy 168 (2025) 127631

Contents lists available at ScienceDirect

European Journal of Agronomy

journal homepage: www.elsevier.com/locate/eja

CO-ResNetRS50-SSL: Enhanced convolution and semi-supervised learning for accurate rice growth stage recognition in complex field conditions

Changqing Yan [a], Guangpeng Yang [a], Zeyun Liang [a], Han Cheng [a], Genghong Wu [b], Amit Kumar Srivastava [c], Qiang Yu [b], Gang Zhao [b,*]

[a] College of Intelligent Equipment, Shandong University of Science and Technology, Taian 271019, China
[b] College of Soil and Water Conservation Science and Engineering, Northwest A&F University, Yangling 712100, China
[c] Leibniz Centre for Agricultural and Landscape Research (ZALF), Multi-Scale Modelling and Forecasting, Eberswalder Strasse 84, Müncheberg 15374, Germany

ARTICLE INFO

Keywords:
Rice
Growth stage identification
ResNetRS50
Omni-Dimensional Dynamic Convolution
Coordinate Attention
Confidence threshold

ABSTRACT

Effective crop management decisions, such as fertilization, irrigation, and crop protection, are closely tied to the crop growth stages. Precise identification of development stages is essential to optimize management practices in line with crop needs. While deep learning has shown promise in identifying growth stages, existing models often face challenges due to limited data availability and reduced accuracy in complex field conditions. To overcome these limitations, this study proposes a semi-supervised image classification method built on an enhanced ResNetRS50 architecture, named CO-ResNetRS50-SSL. This model leverages ResNetRS50 as its backbone, integrating Coordinate Attention (CA) for improved positional feature extraction and Omni-Dimensional Dynamic Convolution (ODConv) to enhance the adaptability of convolutional kernels to varying targets. Additionally, a semi-supervised learning framework is employed to boost generalization while minimizing dependence on labeled data. Ablation experiments show that semi-supervised learning boosted ResNetRS50's accuracy from 88.58 % to 89.36 %. Adding Coordinate Attention further increased accuracy to 89.89 %, while incorporating ODConv in the final CO-ResNetRS50-SSL model achieved 90.38 % accuracy, 90.59 % precision, and 90.19 % F1 score (with 65.38 M parameters). Comparisons reveal that CO-ResNetRS50-SSL outperforms state-of-the-art models (FasterNet-T1, ShuffleNetV2, Swin Transformer, Vision Transformer, ConvNeXt-base) with highly significant differences ($p < 0.001$) and delivers robust performance across rice growth stages, with an optimal trade-off at 224 × 224 resolution. CO-ResNetRS50-SSL can accurately detect rice growth stages with limited labeled data, and its improvements in accuracy and generalization are expected to enhance decision-making in precision agriculture, optimizing resource allocation, reducing inputs, and advancing progress in the field of digital agriculture. Future work will focus on improving efficiency in utilizing unlabeled data, ensuring more balanced performance across different growth stages, and enhancing the model's adaptability to other crops and more complex agricultural scenarios.

1. Introduction

Different crop growth stages, such as germination, vegetative growth, flowering, and grain filling, each require specific amounts of nutrients, water, growth regulators, and protection from pests and diseases (Ravlić et al., 2022; Sun et al., 2025; Zhao et al., 2024b). Consequently, farmers often schedule their management practices to align with these developmental stages (Nyéki and Neményi, 2022; Yue et al., 2020). Precise identification of these stages is crucial for implementing appropriate agricultural interventions, optimizing resource use, and reducing environmental impacts (Coleman et al., 2024; Roy and Bhaduri, 2022). However, accurately distinguishing between growth stages can be challenging due to the subtle similarities in plant appearance and physiological processes across stages (Cortinas et al., 2023; de de de Castro Pereira et al., 2022).

Machine learning (ML)-based computer vision significantly advances the automation of crop trait detection, revolutionizing traditional crop scouting methods (Hu et al., 2023; Liu and Xu, 2023; Tian et al., 2024). Traditional machine learning methods, such as random forest, k-nearest neighbors, Gaussian naïve Bayes, support vector machine, and logistic

* Corresponding author.
E-mail address: gang.zhao@nwafu.edu.cn (G. Zhao).

https://doi.org/10.1016/j.eja.2025.127631
Received 5 February 2025; Received in revised form 25 March 2025; Accepted 27 March 2025

regression, have been used to estimate phenological stages in rice breeding. For example, ensemble models have achieved an accuracy of 93 % on a validation dataset (Ge et al., 2021). Additionally, Zhang et al. (2021) combined support vector machines with principal component analysis for rice tiller period identification, achieving an accuracy of 97.76 %, while Sheng et al. (2022) developed a random forest-based model that identified rice growth stages with 98.772 % accuracy. Despite these successes, such approaches typically perform well only on images with simple backgrounds and often struggle with high-dimensional data and complex nonlinear relationships. Additionally, they heavily depend on manually designed features, which can limit further improvements in accuracy (Janiesch et al., 2021).

The emergence of deep learning techniques has significantly enhanced crop trait identification, particularly in automating the detection of growth stages (Aich et al., 2018; Ferentinos, 2018; Wang et al., 2023; Yan et al., 2024; Yu et al., 2023). For instance, deep learning models like Convolutional Neural Networks (CNNs) can automatically learn abstract features from images, improving classification accuracy without the need for manually designed features (Alabsi et al., 2023). Notable examples include Xiao et al. (2022), who developed a MobileNets-based model to classify Phalaenopsis orchid growth stages with 98.9 % accuracy, and Tan et al. (2023), who introduced RiceRes2Net for detecting rice panicles and growth stages with high accuracy in complex field environments. Schieck et al. (2023) employed ResNet, DenseNet, and InceptionV3 to differentiate developmental stages of grapes at the microscopic level, with ResNet achieving the highest classification accuracy, yielding an average of 88.1 %. While deep learning-based methods have proven effective in crop stage identification, their performance is highly dependent on the availability of large-scale and accurately labeled image datasets. The process of manually labeling these datasets is known for being error-prone, cumbersome, expensive, and time-consuming (Deng et al., 2025; Janiesch et al., 2021). Furthermore, these models often excel in identifying distinct growth stages with clear traits, such as booting, heading, and grain filling in rice (Tan et al., 2023), but their applicability is limited across the entire crop growth cycle. This gap highlights the need for enhanced model training that encompasses a broader range of growth stages to support comprehensive crop management throughout the entire cultivation period.

In recent years, semi-supervised learning (SSL) has garnered significant attention for its ability to effectively combine both labeled and unlabeled data to enhance model performance. For instance, Amorim et al. (2019) demonstrated the effectiveness of semi-supervised methods in deep learning, highlighting that the integration of SSL with deep learning can achieve higher classification accuracy. Khan et al. (2021) proposed an optimized SSL approach for classifying crops and weeds during early growth stages. Their experiments indicated that their method outperformed traditional supervised learning techniques under conditions with a higher proportion of unlabeled data. Li and Chao (2021) further introduced a semi-supervised few-shot learning method for plant leaf disease identification, verifying that the superiority of their method over other related techniques when labeled training data was limited. Benchallal et al. (2024) proposed a novel deep learning architecture based on a semi-supervised learning paradigm, comprising a modern ConvNeXt-Base encoder and a carefully designed decoder for accurate weed species identification. These studies underscore the benefits of SSL when labeled data are limited, making it particularly well-suited for comprehensive crop growth stage recognition. However, integrating advanced deep learning models with SSL under complex field conditions remains underexplored.

To address these challenges, we hypothesize that introducing Coordinate Attention (CA) (Jia et al., 2024) for enhanced positional feature extraction and Omni-Dimensional Dynamic Convolution (ODConv) (Guo et al., 2023) for adaptive feature extraction into the ResNetRS50 architecture—combined with SSL—will enable efficient and accurate rice growth stage recognition under complex field conditions.

In this paper, we propose CO-ResNetRS50-SSL, a deep learning method that integrates CA and ODConv into an improved ResNetRS50 network with SSL. Our main contributions are as follows:

1) A comprehensive dataset was constructed, fully considering the characteristics of rice plants at different growth stages, spanning from seeding to mature plants.
2) A semi-supervised learning method, CO-ResNetRS50-SSL, based on the ResNetRS50 architecture improved with CA and ODConv, was proposed, achieving high-accuracy recognition of key rice growth stages.
3) Extensive experiments including ablation and comparative experiments validated the high performance of our method. The varying performance on different growth stage identification is clarified, providing practical guidance for precise monitoring and management in smart agriculture.

2. Materials and methods

2.1. Overall workflow

Fig. 1 shows a three-part workflow encompassing data, model, and experiment. In the data phase, 21,619 rice images were collected, a portion of which was labeled according to the BBCH scale (Lancashire et al., 1991) and split into training, validation, and test sets, while 6486 unlabeled images were reserved for semi-supervised learning. During the model phase, ResNetRS50 was trained on labeled data to establish a baseline, then enhanced by incorporating SSL (yielding ResNetRS50-SSL), Coordinate Attention (CA) for positional feature extraction (C-ResNetRS50-SSL), and Omni-Dimensional Dynamic Convolution (ODConv) for adaptive kernels (CO-ResNetRS50-SSL). Finally, in the experiment phase, ablation studies isolated the individual contributions of SSL, CA, and ODConv, comparative analyses measured performance against contemporary architectures, and different confidence thresholds (CTs), BBCH-stage evaluations, and image resolutions were tested to balance predictive accuracy with computational cost.

2.2. Dataset collection and pre-processing

2.2.1. Image data collection

The dataset was collected from rice fields across China by Northwest A&F University in a joint project with BASF Digital Farming GmbH (Zhao et al., 2024b) (Fig. 2). These fields were selected to assess the progression of rice diseases under natural conditions, without the application of fungicides. To ensure that the dataset accurately represents the full rice crop life cycle, data collection and annotation were conducted by experienced technicians within a WeChat Mini application, named “NongQingZhuShou”. In the application, photographs were systematically taken every 3–4 days within the experimental fields using smartphones. The dataset encompasses images under a wide range of environmental conditions, including various weather scenarios and soil types, thereby enhancing the robustness and generalizability of the trained models. In total, 21,619 images were captured across the rice fields, as illustrated in Fig. 2.

2.2.2. Image data pre-processing

The images were classified and labeled according to the BBCH scale (Lancashire et al., 1991), a standardized scale widely used for describing the phenological development of plants. The principal BBCH stages include BBCH 1 (Leaf development), BBCH 2 (Tillering), BBCH 3 (Stem elongation), BBCH 4 (Booting), BBCH 5 (Heading), BBCH 7 (Development of fruit), and BBCH 8 (Ripening) (Wang et al., 2022), as shown in Fig. 3(a), were used to label the images. BBCH 6, which corresponds to flowering, is not individually represented in this dataset, as it often occurs concurrently with BBCH 5 in rice, making it challenging to differentiate them distinctly in field conditions. Of all the 21,619

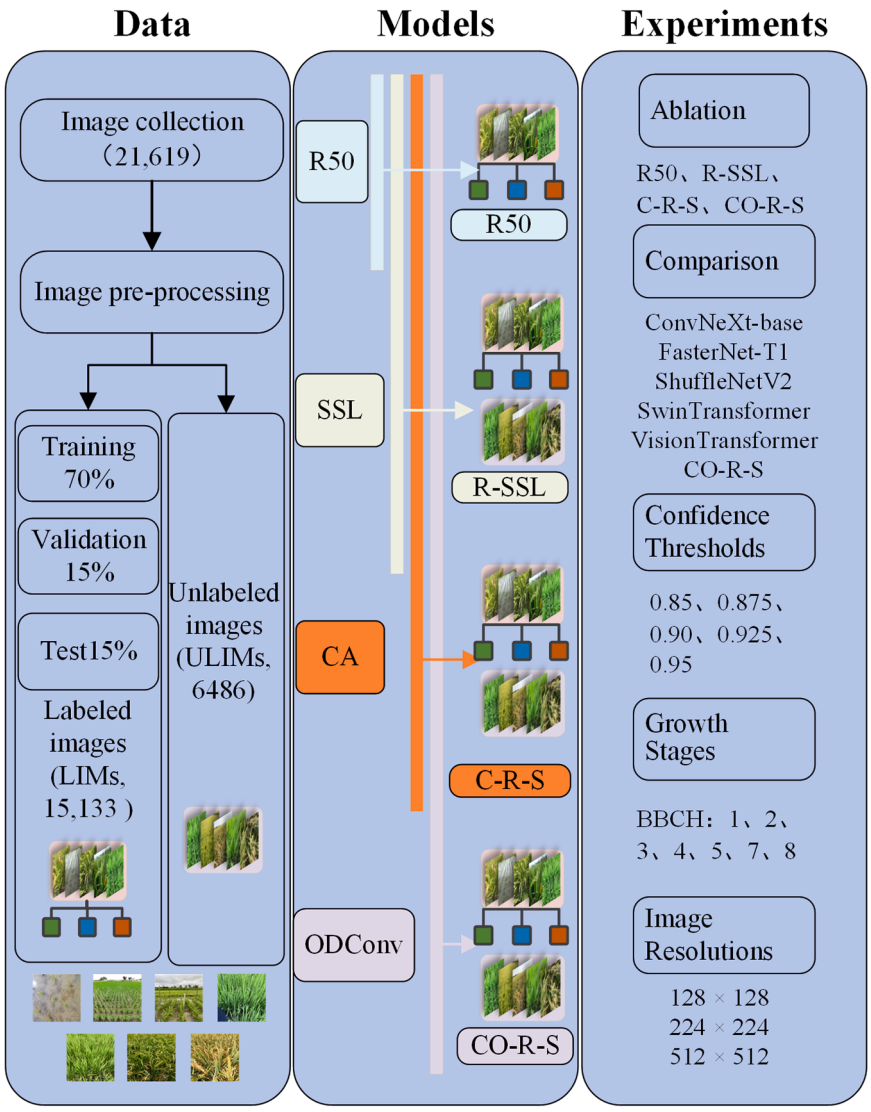

Fig. 1. Overall workflow for data preparation, model improvement, and experiments. SSL represents semi-supervised learning, CA for Coordinate Attention, ODConv for Omni-Dimensional Dynamic Convolution. R50 represents ResNetRS50, R-SSL for ResNetRS50-SSL, C-R-S for C-ResNetRS50-SSL, and CO-R-S for Co-ResNetRS50-SSL. The abbreviations of the models are explained on the top of the data section.

images, we labeled 15,133 images, while 6486 images remained unlabeled, with a ratio of 7:3.

Due to the high resolution of the images captured by smartphones, these images could lead to insufficient GPU memory and slow processing speed when used directly. To address this, all images were resized to 224 × 224 resolution, ensuring they meet the basic requirements for model training while accelerating the training speed.

The resized dataset was then randomly split into training, validation, and testing sets for each BBCH scale in a ratio of 7:1.5:1.5. Specifically, the training set consisted of 10,593 images, while the validation and testing dataset each contained 2270 images, as detailed in Fig. 3(b). This dataset is generally well-balanced across various growth stages. Nevertheless, there is some variability in the number of images due to limitations in data collection, such as a smaller number of BBCH5 images compared to other stages.

2.3. ResNetRS50 as a baseline model

In this study, we selected ResNetRS50 as the backbone architecture. ResNetRS, an enhanced variant of the original ResNet (Bello et al., 2021), further improves performance through refined training strategies and architectural optimizations, as evidenced by its superior results in large-scale benchmarks. Among the ResNetRS series, deeper models such as ResNetRS101 and ResNetRS152, with parameter counts of 93.98 M and 131.69 M respectively, excel in feature extraction due to their increased depth and capacity (Table 1). However, their high computational complexity and resource demands make them less practical for scenarios requiring efficiency. In contrast, ResNetRS50, with a significantly lower parameter count of 48.19 M, achieves a favorable balance between computational efficiency and performance, making it an optimal choice for this study.

The architecture of ResNetRS50, illustrated in Fig. 4, is composed of three primary components: a StemBlock, multiple ResidualBlock groups, and a fully connected (FC) layer. The StemBlock serves as the initial feature extractor, reducing spatial dimensions while increasing channel depth. This is followed by four distinct ResidualBlock group modules, each configured with a varying number of ResidualBlocks—specifically, 3, 4, 6, and 3 blocks, respectively. These groups are designed to progressively extract hierarchical features, with each group increasing the network's depth and complexity while maintaining computational efficiency through skip connections, which mitigate vanishing gradients and enable stable training. The final FC layer acts as the classifier,

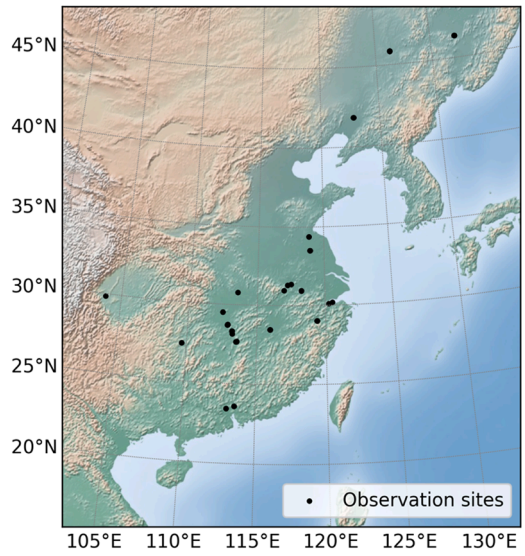

Fig. 2. Map of experimental rice field locations where photos were captured; each point represents a unique observation site of experimental rice field for the image dataset collection.

mapping the extracted features to the output space.

During the training of ResNetRS50, several data augmentation techniques were employed to enhance the model's robustness and generalization capabilities. These techniques included random cropping, horizontal flipping, and channel normalization. Random cropping was applied to reduce potential biases in the dataset and encourage the model to focus on spatially invariant features. Horizontal flipping further diversified the training data by introducing mirrored versions of the images, improving the model's ability to handle variations in orientation. For normalization, each pixel was standardized using pre-computed mean and standard deviation values for the Red, Green, and Blue (RGB) channels. Specifically, the mean and standard deviation values were 0.5071 ± 0.2673 for the R channel, 0.4865 ± 0.2564 for the G channel, and 0.4409 ± 0.2762 for the B channel. This normalization process ensured that the input data was centered and scaled, reducing the risk of bias and accelerating convergence during training. The final trained ResNetRS50 model, optimized through these techniques, served as the baseline for further enhancements using SSL, aiming to improve its performance and adaptability to downstream tasks.

2.4. Semi-supervised learning for enhancing ResNetRS50

Although ResNetRS50 demonstrated strong performance as a baseline model, its training process is heavily dependent on a large volume of labeled data, which can be expensive and labor-intensive to acquire in real-world scenarios. To address this limitation and harness the potential of abundant unlabeled data, we developed a semi-supervised learning framework, as depicted in Fig. 5. The SSL approach begins by utilizing the pre-trained ResNetRS50 model to predict growth stages for unlabeled images, generating pseudo-labeled images (PLIMs). However, low-confidence pseudo-labels may introduce inaccuracies, potentially leading the model to learn incorrect patterns. To mitigate this risk, a confidence threshold (CT) was implemented to filter out unreliable predictions. Only PLIMs with confidence scores exceeding the threshold were retained as selected pseudo-labeled images (SPLIMs), while those below the threshold were discarded. The SPLIMs were then combined with the original labeled image dataset (LIMs) to form an augmented dataset (LIMs + SPLIMs), which was used to retrain the ResNetRS50 model.

2.5. Coordinate attention and omni-dimensional dynamic convolution enhanced ResNetRS50 model

Despite the high performance of ResNetRS50 and the deep features extraction ability, but its accuracy drops when the background is cluttered (Lyu et al., 2023). To enable growth stage identification in real rice fields with complex backgrounds, we propose the CO-ResNetRS50 model—an enhanced version of ResNetRS50 that integrates Coordinate Attention (CA) and Omni-Dimensional Dynamic Convolution (ODConv) modules. Its architecture is depicted in Fig. 6. ResidualBlock is the basic unit of the network, determining the overall performance of the network to a large extent. Therefore, the network was improved by enhancing the ResidualBlocks with the incorporation of CA and ODConv, resulting in the COResidualBlock. Firstly, CA was introduced after the third BatchNormReLU layer and before the SE module in the COResidualBlock, with the reduction parameter set to 32 (Fig. 6, c). This modification aimed to enhance ResNetRS50 network's capability to extract positional information while avoiding significant computational overhead. Secondly, the convolution following the downsampling and preceding the first BatchNormReLU layer was replaced with ODConv in the COResidualBlock. In this configuration, the convolutional layer uses 4 kernels, each with a size of 1 and a stride of 1 (Fig. 6, c). This substitution was intended to improve the convolutional kernels' ability to capture target features, enhance the network's sensitivity to feature extraction, and minimize the increase in model computational complexity as much as possible.

2.5.1. Coordinate attention

The field environment is characterized by its complexity and diversity, with background elements such as weeds, soil, and varying lighting conditions introducing significant noise. This makes it challenging for models to rely solely on global semantic features for accurate judgments. Therefore, it is essential for models to identify and localize fine-grained features in the target regions, such as the position of specific leaves, the distribution of panicles, or the arrangement of plants. Capturing detailed positional information allows models to distinguish target features from the complex background, enhancing their ability to differentiate between different growth stages. While ResNetRS50 excels in extracting high-level semantic features, it tends to overlook crucial positional information and struggles to capture long-distance dependencies between features (Panda et al., 2022; Guo et al., 2022). This limitation is especially evident in the complex field conditions of rice crop where both spatial and semantic details are crucial for accurate stage identification. To improve performance in such tasks, we enhanced ResNetRS50's ability to extract and interpret positional information, thereby enabling more accurate differentiation between growth stages.

Incorporating mechanisms like Coordinate Attention (CA) can address this limitation by embedding spatial dependencies into the model's feature extraction process (Hou et al., 2021). CA enhances the model's ability to focus on both specific image regions and their relative positions, which is particularly valuable for rice growth stage recognition in complex scenes. By horizontally and vertically integrating feature information through two 1D global pooling operations, CA encodes feature maps into two attention maps, allowing the model to capture long-distance dependencies. This structure strengthens the model's understanding of spatial relationships, improving its overall performance in tasks that require detailed spatial feature extraction. As illustrated in Fig. 7, CA consists of four main components that work together to guide the model's attention towards relevant positional information. This mechanism enables the model to concentrate on key growth stage features of rice plants while minimizing the influence of irrelevant background details. The specific computation steps and corresponding operations are described below (Hou et al., 2021).

First, it performs average pooling on the feature map separately along the vertical and horizontal directions. This operation ensures that each output feature point contains information specific to either the

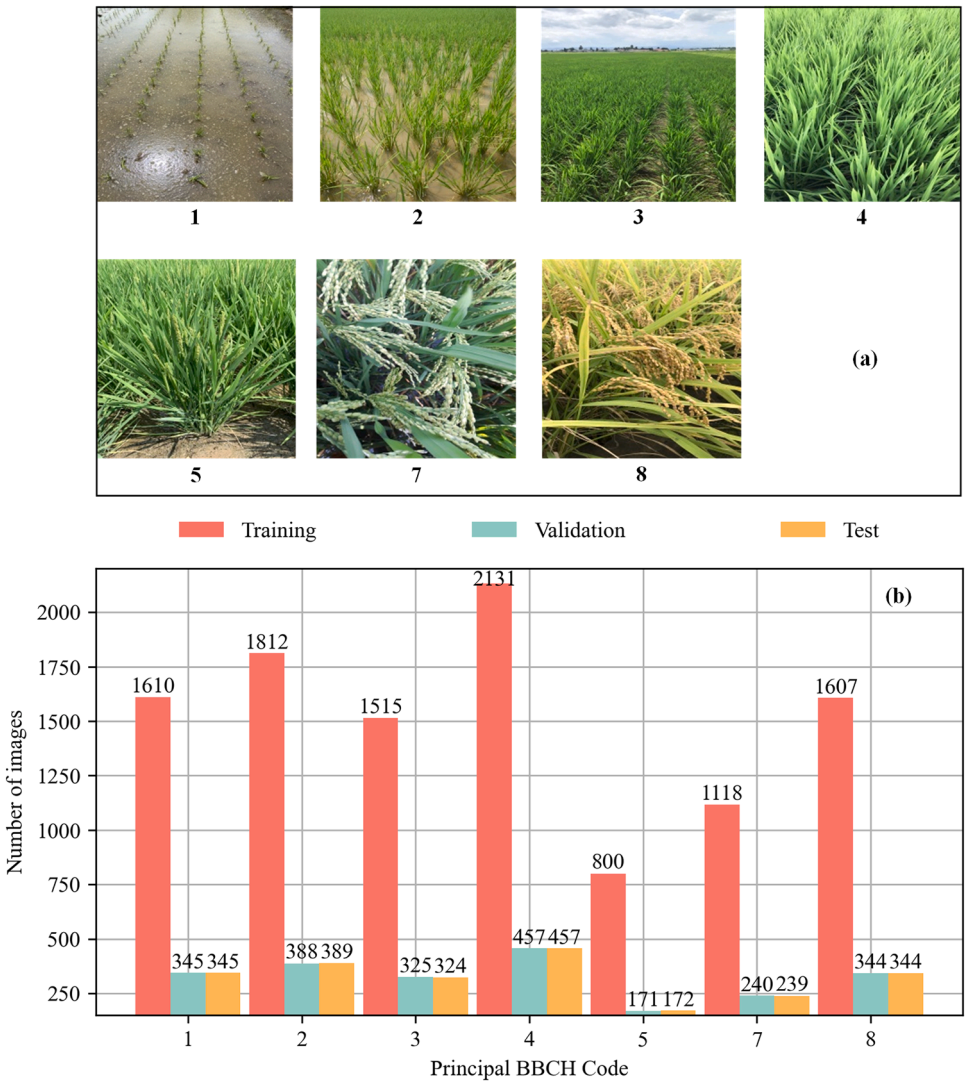

Fig. 3. Representative image examples of rice growth stages and morphological features captured during field measurements, along with the corresponding distribution of the images across the training, validation, and test datasets. (a) Examples showcasing various growth stages and rice morphologies. (b) Distribution of rice images by growth stage (in BBCH code) in the training, validation, and test datasets.

Table 1
Comparison of ResNetRS Series Models (Bello et al., 2021) in Terms of Parameter Count, Latencies on Tesla V100 GPUs (V100 Lat), Latencies on TPUv3 (TPU Lat), and Top-1 Accuracy. Results are based on training the models on the ImageNet dataset using TensorFlow 1.[11]

Model	Parameter (M)	V100 Lat (s)	TPU Lat (ms)	Top-1 Accuracy (%)
ResNetRS50	48.19	0.31	70	78.8
ResNetRS101	93.98	0.70	170	81.2
ResNetRS152	131.69	1.48	320	82.2

vertical or horizontal direction, thereby capturing the structural patterns present in complex field images more effectively. By transforming the feature channels in both the vertical and horizontal directions, the network can more precisely localize the target of interest. The operations for average pooling in the vertical and horizontal directions are shown in Eq. 1 and Eq. 2.

Eq. 1 represents the sum of all pixel values in the width direction at a specific height h in the feature map x_C, divided by the width W, resulting in the average value $\alpha_C^h(h)$ at that height.

$$\alpha_C^h(h) = \frac{1}{w} \sum_{0 \le j < w} x_c(h, i) \tag{1}$$

where $\alpha_C^h(h)$ represents the pooling result in the vertical (height) direction of the feature map, indicating the mean pixel value at height h across all width directions.

Eq. 2 represents the sum of all pixel values in the height direction at a specific width w in the feature map x_C, divided by the height H, resulting in the average value $\alpha_C^w(w)$ at that width.

$$\alpha_C^w(w) = \frac{1}{H} \sum_{0 \le j < H} x_c(j, w) \tag{2}$$

where $\alpha_C^w(w)$ represents the pooling result in the horizontal (width) direction of the feature map, indicating the mean pixel value across all height directions at the width w.

Afterwards, the feature maps after channel pooling are concatenated to obtain a new feature map, which contains both vertical and horizontal global information. This global information enhances the model's ability to better understand the growth characteristics of the rice plants. A 1×1 convolution operation is then applied to the concatenated feature map

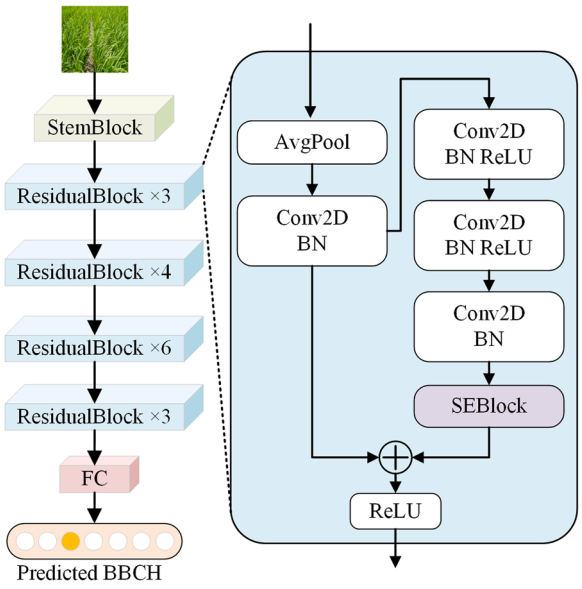

Fig. 4. Overview of the ResNetRS50 architecture, which begins with a Stem Block for initial feature extraction, followed by multiple Residual Blocks for deep feature learning, and concludes with a FC (Fully Connected) layer for final classification.

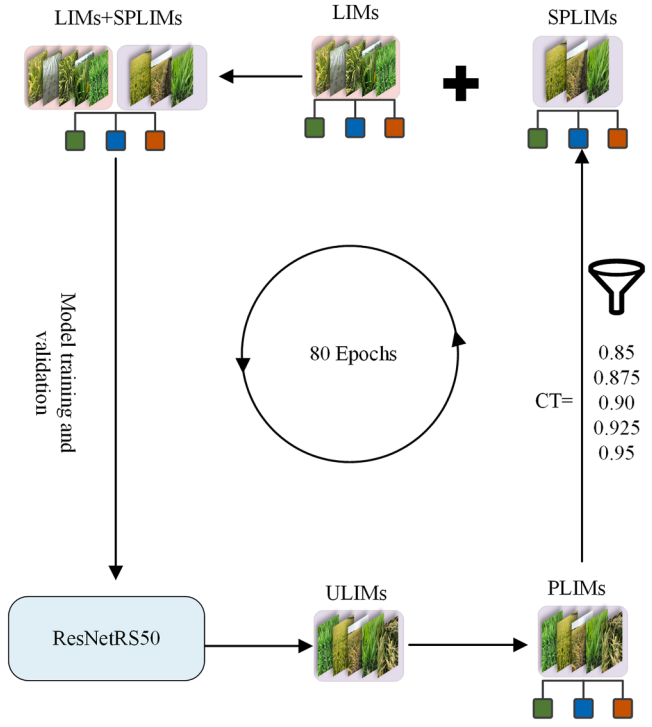

Fig. 5. Overview of the proposed semi-supervised learning pipeline for rice growth stage recognition. ULIMs (Unlabeled images) are used to generate PLIMs (pseudo-labeled images), which are then filtered by a CT (confidence threshold) to produce SPLIMs (selected pseudo-labeled images). These SPLIMs are combined with LIMs (labeled images) for training. For each CT value (0.85, 0.875, 0.90, 0.925, 0.95), ResNetRS50 is trained for 80 epochs.

to integrate the information and reduce the number of channels. Following this, the features are normalized through a batch normalization layer to accelerate the training process and enhance the stability of the model. Finally, a nonlinear activation function is applied to the convolution result to obtain the intermediate feature map, as shown in Eq. 3.

$$f = \sigma\left(T_1\left[\alpha_C^h, \alpha_C^w\right]\right) \tag{3}$$

where $[\alpha_C^h, \alpha_C^w]$ represents the concatenation of α_C^h and α_C^w. T_1 denotes a 1×1 convolution, σ is the non-linear activation, and f represents the intermediate feature map, where $f \in R^{C/r\times(W+H)}$. The reduction ratio r controls the parameter size, and is set to the value of $r = 32$, a value chosen based on previous research (Hou et al., 2021). Reducing r increases the model's parameter count. Therefore, the default setting strikes an optimal balance between performance and complexity.

Then, a 1×1 convolution operation is applied separately to the vertical feature map and the horizontal feature map, which are extracted from the intermediate feature map, to integrate the information and generate feature weights. This design improves the model's ability to capture directional attention, enabling it to differentiate between the vertical growth patterns of rice panicles from the horizontal expansion of leaves. Subsequently, a Sigmoid activation function is applied to the convolution results to obtain the vertical feature weights and horizontal feature weights. The specific operations are shown in Eq. 4 and Eq. 5 .

$$\beta_C^h = \delta\left(T_h\left(f^h\right)\right) \tag{4}$$

where f^h represents the vertical intermediate feature map with $f^h \in R^{C/r\times H}$. T_h is a vertical 1×1 convolution, δ is the Sigmoid activation, β_C^h denotes the vertical feature weight with $\beta_C^h \in R^{C\times H\times 1}$.

$$\beta_C^w = \delta(T_w(f^w)) \tag{5}$$

where f^w represents the horizontal intermediate feature map with $f^w \in R^{C/r\times W}$. T_w is a horizontal 1×1 convolution, β_C^w represents the horizontal feature weight with $\beta_C^w \in R^{C\times 1\times W}$.

At the end, vertical and horizontal attention weights are applied to each pixel of the original input feature map through a broadcasting mechanism, thereby generating an output feature map that contains positional information. This can be obtained by Eq. 6.

$$y_C(i,j) = x_C(i,j) \times \beta_C^h(i) \times \beta_C^w(j) \tag{6}$$

where $y_C(i,j)$ represents the pixel value at position (i,j) in the output feature map, while $x_C(i,j)$ represents the pixel value at position (i,j) in the original input feature map. $\beta_C^h(i)$ denotes the vertical feature weight at row i, and $\beta_C^w(j)$ represents the horizontal feature weight at column j.

Through these steps, the model is able to focus on regions with important spatial relationships, avoiding the loss of positional information caused by traditional 2D global pooling, thereby enhancing ResNetRS50's performance in tasks with complex background features.

2.5.2. *Omni-dimensional dynamic convolution*

The traditional convolution operation employs fixed convolution kernels, and these static weights cannot adjust based on image content, making it difficult to adapt to the complex and dynamic background of field environments, such as interference from weeds, soil, and shadows. Therefore, Omni-Dimensional Dynamic Convolution (ODConv) (Li et al., 2022) was introduced to replace the standard convolutions in ResNetRS50, enhancing the convolution kernels' adaptive ability to rice crop features under complex field conditions.

ODConv utilizes a novel multi-dimensional attention mechanism to compute four types of attention, α_{si}, α_{ci}, α_{fi}, and α_{wi}, in a parallel manner along all four dimensions of the kernel space (spatial dimension, input channel dimension, output channel dimension, and kernel dimension of the convolutional kernel space)(Fig. 8). Specifically, a global average

[1] V100 Lat represents Latencies on Tesla V100 GPUs, and TPU Lat represents Latencies on TPUv3.

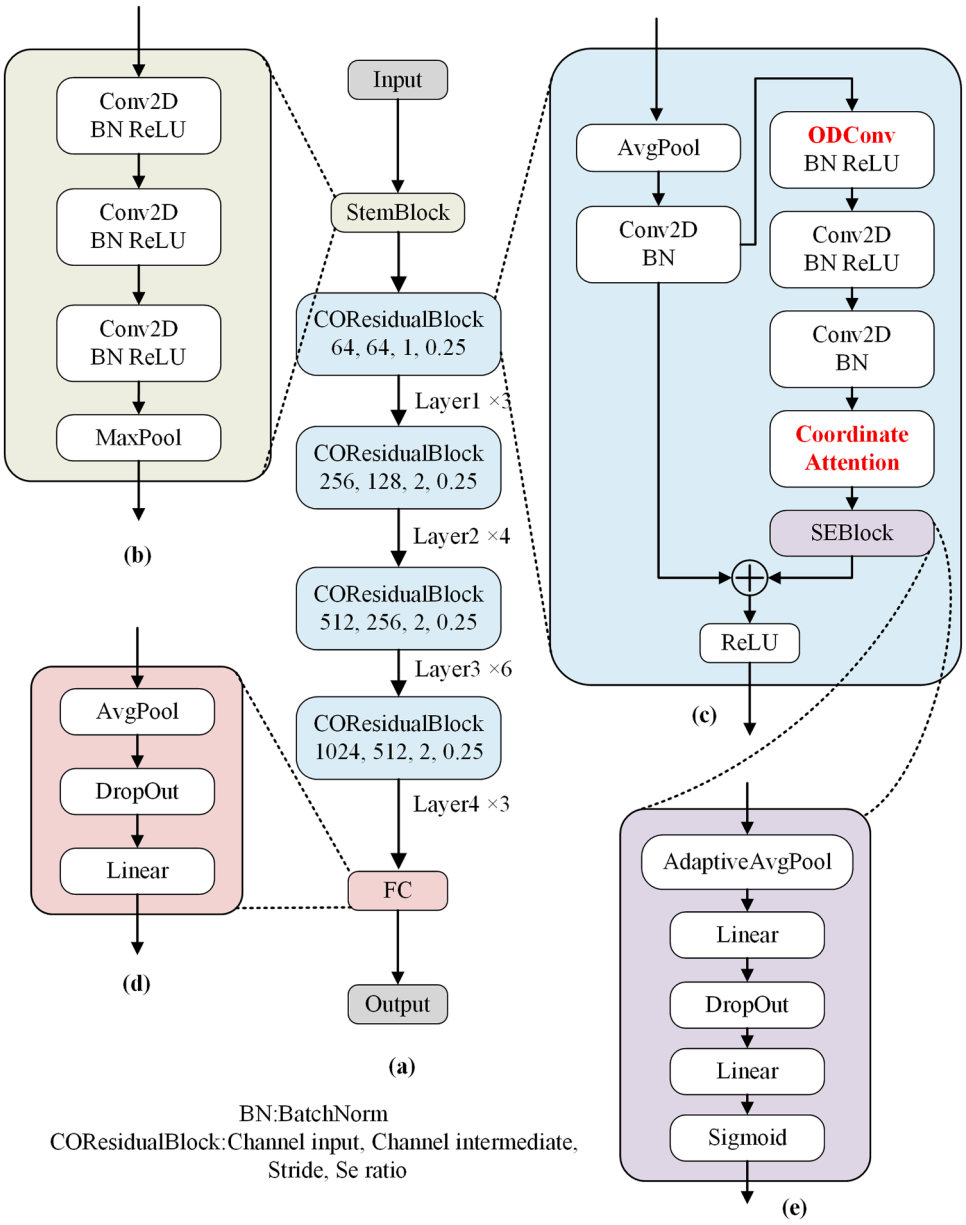

Fig. 6. Network architecture of the proposed CO-ResNetRS50, highlighting ODConv (Omni-Dimensional Dynamic Convolution) and Coordinate Attention (in red) as the key improvement modules. (a) The overall pipeline consists of a StemBlock, multiple COResidualBlocks, and a FC (Fully Connected) layer. (b) The StemBlock includes Conv2D, BatchNorm, ReLU, and MaxPool layers. (c) Each COResidualBlock integrates AvgPool, Conv2D, ODConv, BatchNorm, ReLU, Coordinate Attention, and an SEBlock. (d) The FC module is composed of AvgPool, DropOut, and Linear layers. (e) The SEBlock includes AdaptiveAvgPool, Linear, DropOut, and Sigmoid layers.

pooling (GAP) operation is first applied to the input feature map, to compress the original input features x into a feature vector of length c_{in}. This operation effectively captures the global information of rice images along the channel dimension. Subsequently, a fully connected (FC) layer maps the compressed feature vector to a lower-dimensional space with a reduction ratio of 1/16. This mapping process helps reduce computational complexity while focusing on retaining the critical growth stage features of rice, though it may increase some computational cost and memory usage in large-scale datasets or complex scenarios. An ReLU activation function follows the FC layer, introducing non-linearity and enabling the model to flexibly represent the distinctive features of different rice growth stages. Then, four branches calculate attention scalars for each dimension separately. Each branch contains an FC layer and an activation function (Sigmoid or Softmax) to generate normalized attention values. These four attentions, spatial-level, input channel-level, output channel-level, and kernel-level attention, are sequentially applied to the convolution kernel W_i. By dynamically adjusting the weights at each level, the convolution operation becomes adaptive to all spatial positions, input channels, output channels, and kernels of the input x. At the spatial-level attention stage, the model focuses on key areas of the feature map, such as significant regions of leaves and panicles, effectively capturing subtle differences between similar growth stages. In input channel-level attention stage, weights are dynamically adjusted based on the features of the input channels, enhancing the extraction of relevant features from the target area while minimizing the interference from irrelevant information. Similarly, in output channel-level attention stage, the model adaptively modifies weights to emphasize task-relevant features from the output channels. In kernel-level attention, the weights of the convolution kernels are adjusted dynamically based on the input features, improving the model's ability

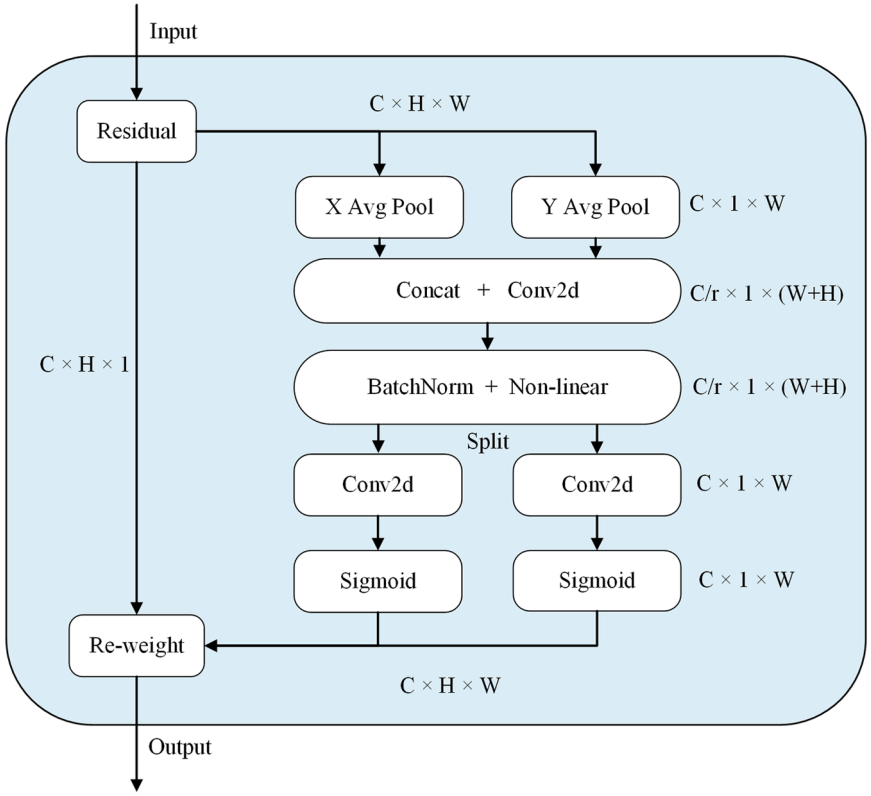

Fig. 7. Diagram of the Coordinate Attention module. The input feature map undergoes separate X Avg Pool and Y Avg Pool (X and Y average pooling), Concat (concatenation), and then passes through Conv2d (convolution), BatchNorm (batch normalization), Non-linear (non-linear activation), Conv2d (convolution), and Sigmoid (sigmoid gating). The resulting attention maps re-weight the original input (Residual), enhancing both spatial and channel-wise feature representation.

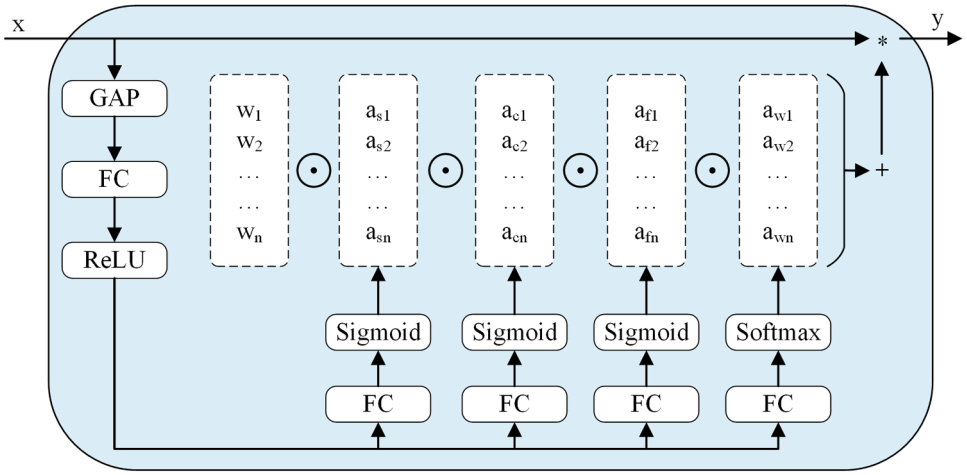

Fig. 8. Schematic of the ODConv (Omni-Dimensional Dynamic Convolution). The structure includes GAP (Global Average Pooling), FC (Fully Connected) layers, ReLU, Sigmoid, and Softmax operations, which collectively compute four attention coefficients (α_{si}, α_{ci}, α_{fi}, and α_{wi}) for the kernel weights (W_i). This multidimensional attention mechanism spans all four dimensions of the kernel space, enabling more adaptive and expressive feature learning.

to focus on essential areas. For instance, kernel-level attention can highlight the edge details of rice panicles while suppressing background noise. Finally, the adjusted convolution kernels are applied to the input features, generating adaptive dynamic convolution results. This convolution operation can automatically adjust the size and shape of the convolution kernels based on the image content, enabling more effective extraction of critical features from the image.

ODConv can be calculated in terms of Eq. 7 (Li et al., 2022):

$$y = \left(\alpha_{w1} \odot \alpha_{f1} \odot \alpha_{c1} \odot \alpha_{s1} \odot W_1 + \cdots + \alpha_{wn} \odot \alpha_{fn} \odot \alpha_{cn} \odot \alpha_{sn} \odot W_n\right) * x \tag{7}$$

where x and y represent the input and output features, respectively; $\alpha_{wi} \in R$ denotes the attention scalar for the convolutional kernel W_i; $\alpha_{si} \in R^{k \times k}$, $\alpha_{ci} \in R^{c_{in}}$ and $\alpha_{fi} \in R^{c_{out}}$ denote three newly introduced attentions, which are computed along the spatial dimension, input channel dimension and output channel dimension of the kernel space for the convolutional kernel W_i, respectively; $\odot$ denotes the multiplication op-

erations along different dimensions of the kernel space; * represents the convolution operation.

2.6. Experimental design and setup

2.6.1. Ablation experiments

The first series of experiments aimed to isolate the impact of each proposed enhancement on the ResNetRS50 architecture. Initially, the baseline model was trained in a fully supervised setting using only the labeled images from the dataset, establishing a reference point in terms of accuracy (Eq. 9), precision (Eq. 10), recall (Eq. 11), F1 score (Eq. 12), training time. Next, semi-supervised learning (SSL) was integrated by generating pseudo-labeled images (PLIMs) from unlabeled data. A confidence threshold (CT) mechanism was employed to filter these PLIMs into selected high-quality pseudo-labeled images (SPLIMs). Subsequently, Coordinate Attention (CA) was incorporated into the ResidualBlock to enhance the network's ability to capture spatial and positional cues. Finally, standard convolutional layers within the ResidualBlock were substituted with Omni-Dimensional Dynamic Convolution (ODConv) to enable dynamic kernel adaptation based on image content. Performance evaluations were conducted after the incremental addition of each module, allowing us to assess both the individual contributions and the cumulative effect on model performance.

To evaluate the effectiveness of the performance enhancement modules, a *t*-test was conducted to assess the statistical significance of the performance differences between the CO-ResNetRS50-SSL model and models with progressively integrated enhancement modules. Each model underwent five independent runs. Then t-tests were performed and p-values were calculated for each performance metric to assess the significance of performance differences. If $p \leq 0.001$ (denoted as ***), it indicates a highly significant difference; if $0.001 < p \leq 0.01$ (denoted as **), it indicates a significant difference; if $0.01 < p \leq 0.05$ (denoted as *), it indicates a moderate difference; and if $p > 0.05$ (denoted as ns), it indicates no statistically significant difference.

2.6.2. Comparative analysis with contemporary models

To comprehensively validate the efficacy of our proposed approach, the CO-ResNetRS50-SSL model was rigorously compared against several state-of-the-art deep learning architectures. The comparative models included ConvNeXt-base (Liu et al., 2022), FasterNet-T1 (Chen et al., 2023), ShuffleNetV2 (Ma et al., 2018), Swin Transformer (Liu et al., 2021), and Vision Transformer (Dosovitskiy et al., 2020). Unlike these models—which were trained exclusively on the fully labeled datasets—the CO-ResNetRS50-SSL model was trained using a combination of the fully labeled and unlabeled data, thereby leveraging additional information to enhance performance. For a robust baseline, the standard ResNetRS50 model was also included in the analysis.

T-tests were also employed to assess whether the performance differences between the CO-ResNetRS50-SSL model and the comparative models were statistically significant. For each metric (accuracy, precision, recall, and F1 score), we compared the outcomes from five independent experimental runs of the CO-ResNetRS50-SSL model with those from five runs of each comparative model, ensuring robust and reproducible performance evaluations. The same evaluation methods were employed as in ablation experiments. This approach provides a rigorous framework for comparing the performance of the proposed method with the contemporary models under controlled and replicable conditions.

2.6.3. Optimal confidence threshold selection for semi-supervised learning

To determine an optimal confidence threshold, we evaluated five candidate values (0.85, 0.875, 0.9, 0.925, and 0.95), chosen to balance the trade-off between the quantity and quality of pseudo-labeled data. A threshold below 0.85 risked introducing excessive noise from low-confidence labels, while a threshold above 0.95 could overly restrict the pool of pseudo-labeled data, diminishing the benefits of SSL. Given that the model typically converged within 80 training epochs, the SSL process was iterated for 80 epochs for each threshold. After each training cycle, the model's performance was assessed using a held-out test dataset. This iterative evaluation allowed us to systematically analyze the influence of different confidence thresholds on model accuracy, ultimately identifying the optimal threshold for integrating pseudo-labeled data into the training process. By leveraging SSL, we significantly reduced the reliance on labeled data while maintaining robust model performance.

2.6.4. Performance evaluation across different rice growth stages

We also investigated performance by principal BBCH code to detect class-specific weaknesses or biases. Each of the four main models (ResNetRS50, C-ResNetRS50-SSL, CO-ResNetRS50-SSL, and an additional baseline) was evaluated on a held-out test set, and we generated confusion matrices to visualize inter-class accuracy.

2.6.5. Impact of image resolution on model accuracy

To assess the impact of image resolution on model performance, the dataset was preprocessed at three resolutions (128 × 128, 224 × 224, and 512 × 512) and split into training and test sets at each resolution. Training and inference both used the same resolution in each experiment. All models were trained with identical hyperparameters, data augmentation strategies, and training schedules, ensuring that any variations in performance resulted solely from changes in resolution. Evaluations covered both predictive metrics (accuracy (Eq. 9), precision (Eq. 10), recall (Eq. 11), F1 score (Eq. 12)) and computational metrics (training time and inference latency). This approach enabled a detailed comparison of accuracy–efficiency trade-offs, facilitating the selection of an optimal resolution that balances performance with computational demands.

2.6.6. Computing environment and configuration

Our study was carried out in a Windows 11 environment using an Intel Core i9–14900K CPU, 128 GB RAM, and a GeForce RTX 4090 GPU with 24 GB VRAM. We used PyTorch 2.4.0 and Python 3.10.14 with CUDA 12.5 for deep learning model training, validation, and testing. Models were trained for 80 epochs, starting with a learning rate of 0.01, which was reduced to one-tenth of the current learning rate every 30 epochs. The batch size was set to 32. The Stochastic Gradient Descent (SGD) optimizer, with a momentum of 0.9 and weight decay of 0.0005, was employed. CrossEntropyLoss was used as the loss function (Mao et al., 2023):

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{ic}\log(p_{ic}) \tag{8}$$

where N is the number of samples; C is the number of classes; y_{ic} is the true label of the i-th sample for class c, where $y_{ic} = 1$ if the sample belongs to class c, otherwise $y_{ic} = 0$; and p_{ic} is the predicted probability of the i-th sample for class c.

2.7. Evaluation metrics

To evaluate the model's classification performance on the rice growth stage dataset and compare it with other network models, multiple metrics were employed, including accuracy, precision, recall, F1-score, parameter size and computing time. These metrics were chosen to provide a comprehensive assessment of each model's effectiveness across various aspects of classification, such as the model's ability to correctly identify growth stages, minimize false positives and negatives, and balance precision and recall. The performance was calculated using the standard equations (Eqs. 9–12) for each metric to ensure objective comparisons (Qin et al., 2023).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

$$F1score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{12}$$

where TP denotes true positives; TN is true negatives; FN is false negatives; FP is false positives.

3. Results

3.1. Cumulative enhancements of ResNetRS50 via SSL, CA, and ODConv in ablation study

The performance improvements across the different model variations in Ablation Experiments are summarized in Fig. 9. The baseline ResNetRS50 achieved 88.58 % accuracy, 88.02 % recall, 88.85 % precision, and an 88.35 % F1 score using 48.19 M parameters. When semi-supervised learning was added, accuracy improved to 89.36 % and precision to 89.84 %, with recall and F1 score also increasing slightly while the parameter count remained the same. Adding coordinate attention further enhanced performance to 89.89 % accuracy, 89.32 % recall, 90.02 % precision, and an 89.62 % F1 score, with a minor parameter increase to 50.11 M. Finally, replacing standard convolution with omni-dimensional dynamic convolution yielded the best overall results—90.38 % accuracy, 89.88 % recall, 90.59 % precision, and a 90.19 % F1 score—though the parameter count rose to 65.38 M. Statistical tests confirmed that these improvements are significant, with most metrics showing p-values below 0.001, except for one precision comparison which was significant at $p < 0.05$.

3.2. Performance comparison with state-of-the-art models

Fig. 10 shows that CO-ResNetRS50-SSL outperforms all other models on various metrics, despite having a relatively modest parameter count (65.38 M). Although FasterNet-T1 (6.32 M parameters) and ShuffleNetV2 (1.26 M parameters) are much smaller, they have significantly lower accuracy (88.41 % and 87.26 %) and F1 scores (87.82 % and 86.42 %). Meanwhile, even though Swin Transformer (27.52 M) and Vision Transformer (85.80 M) have more parameters, their accuracy (83.85 % and 83.73 %) and F1 scores (82.90 % and 82.59 %) remain lower. Similarly, ConvexNet-base also achieves much lower performance with the largest parameter size. These results show that CO-ResNetRS50-SSL not only achieves higher accuracy but also makes more efficient use of parameters. Moreover, t-tests confirm that the differences between CO-ResNetRS50-SSL and the other five models are highly significant ($p < 0.001$), underscoring CO-ResNetRS50-SSL's superior performance.

3.3. Impacts of confidence thresholds on model performance

The results of the confidence threshold (CT) experiments on the test set are shown in Fig. 11. When applying SSL with increasing CTs to ResNetRS50, the model's performance first increased, then declined, followed by a subsequent rise, and ultimately declined again. At the CT of 0.85, all performance metrics were improved against the baseline model. Specifically, accuracy was increased from 88.59 % to 89.30 %, a rise of 0.71 percentage points; recall was slightly improved to 88.50 %; precision was significantly increased to 89.45 %; and the F1 score correspondingly rose to 88.93 %. However, when the CT was further increased to 0.875, although precision was 0.20 % higher than that of the baseline model, all other metrics were the lowest. At the CT of 0.90, the model achieved its highest performance, with an accuracy of 89.38 %, representing a 0.79 % improvement over the baseline model. Recall increased to 88.73 %, while precision and F1 score reached

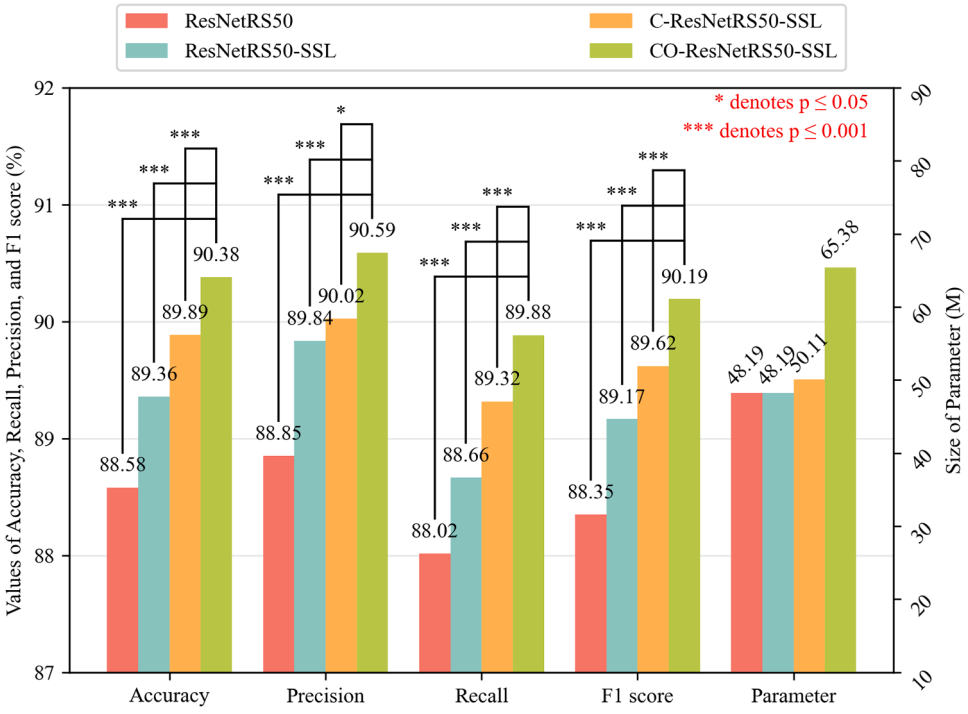

Fig. 9. Ablation study results comparing the baseline ResNetRS50, its semi-supervised extension (denoted as ResNetRS50-SSL), the addition of Coordinate Attention (denoted as C-ResNetRS50-SSL), and the final integration of ODConv (Omni-Dimensional Dynamic Convolution) module (denoted as CO-ResNetRS50-SSL). Bars represent accuracy, recall, precision, F1 score, and parameter count, with $p < 0.05$ and $p < 0.001$ indicating statistically significant improvements over the previous variant.

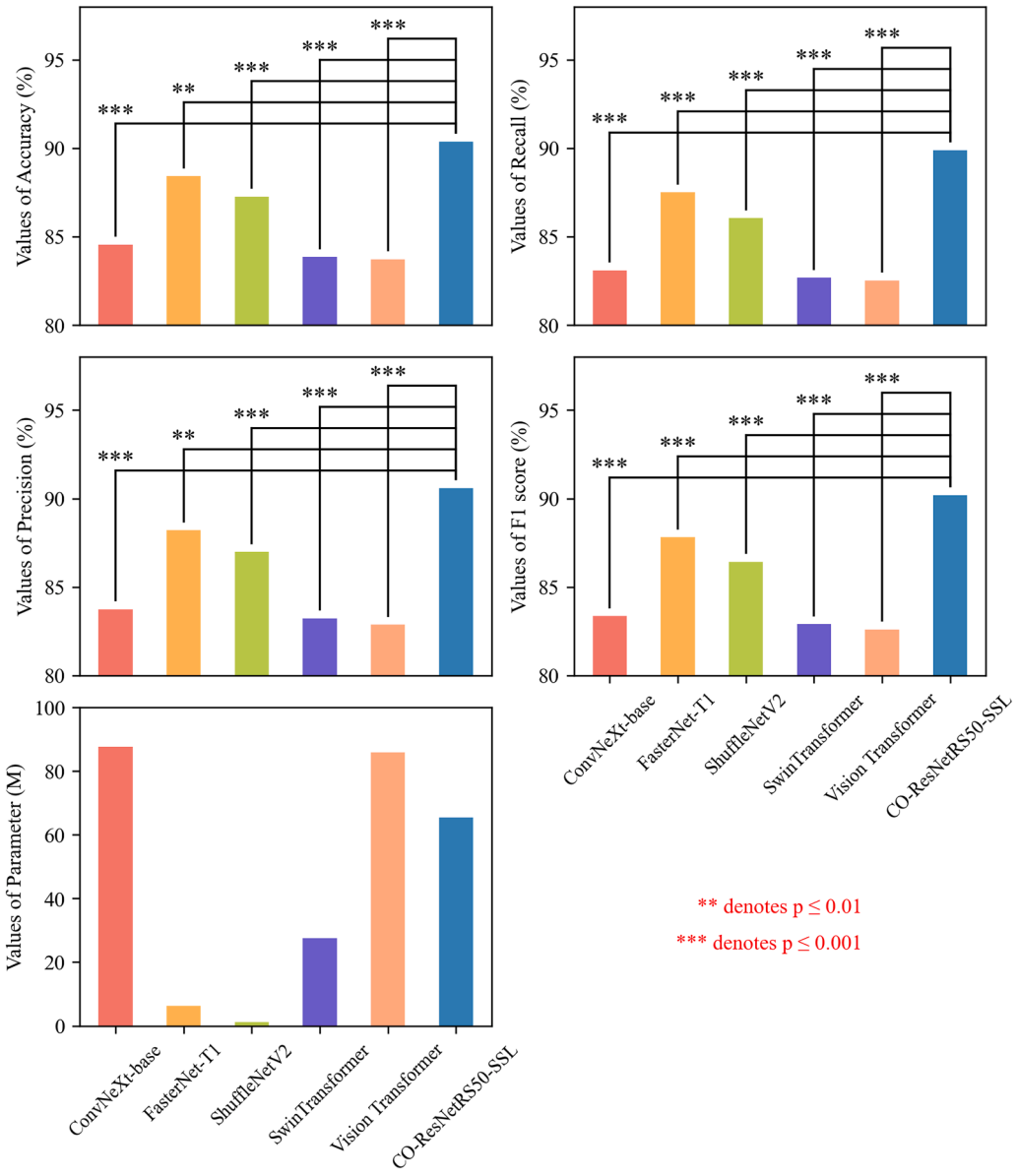

Fig. 10. Performance comparison of our model CO-ResNetRS50-SSL to five State-of-the-Art Models (ConvNeXt-base, FasterNet-T1, ShuffleNetV2, SwinTransformer, Vision Transformer) on the test set, evaluated using metrics including accuracy, recall, precision, F1 score, and parameter count. * * denotes p ≤ 0.01, * ** denotes p ≤ 0.001, with smaller p-value indicating stronger significance in the difference between the model and CO-ResNetRS50-SSL.

89.38 % and 89.17 %, respectively. However, when the CT was raised to 0.95, precision improved further to 89.95 % before declining, while all other metrics showed varying degrees of decline.

Fig. 12 illustrates the changes of selected pseudo-labeled images (SPLIMs) numbers added with the increasing epoch numbers under different CTs in SSL. It is evident that, under the same epoch, the number of added unlabeled samples gradually decreased with the increasing CT. Notably, the number of images added by ResNetRS50-SSL-85 was significantly higher than that of other methods and the number added by ResNetRS50-SSL-925 was the lowest. Further, despite the highest accuracy achieved by the model with CT of 0.90, the number of pseudo-labeled image added by it was moderate. Overall, ResNetRS50 model utilizing SSL was generally found to perform better than the original ResNetRS50 model. At the CT of 0.90, the best overall performance was achieved, achieving the highest accuracy while using significantly fewer unlabeled images than the 0.85 and 0.875 thresholds. Therefore, for subsequent experiments, 0.90 was selected as the CT for SSL.

3.4. Performance at different rice growth stages

The results of different rice growth stages recognition using different models are shown in Fig. 13. In BBCH 1 (leaf development), CO-ResNetRS50-SSL achieved the highest accuracy of 96.52 %, followed by ResNetRS50 and C-ResNetRS50-SSL with slight variations. For BBCH 2 (tillering), all models performed with similar high accuracy, with C-ResNetRS50-SSL slightly outperforming the others. In BBCH 3 (stem elongation), ResNetRS50 had the highest accuracy, while the other models showed lower performance. For BBCH 4 (booting), both CO-ResNetRS50-SSL and C-ResNetRS50-SSL achieved the highest accuracy of 91.90 %. In BBCH 5 (inflorescence emergence), CO-ResNetRS50-SSL also showed the highest accuracy of 84.30 %, while the other models demonstrated lower performance. For BBCH 8 (ripening), CO-ResNetRS50-SSL again performed strongly, although C-ResNetRS50-SSL slightly outperformed it in BBCH 7 (fruit development). Overall, CO-ResNetRS50-SSL generally achieved the highest performance, particularly in the later growth stages (BBCH 1, 4, 5, and 8), while ResNetRS50

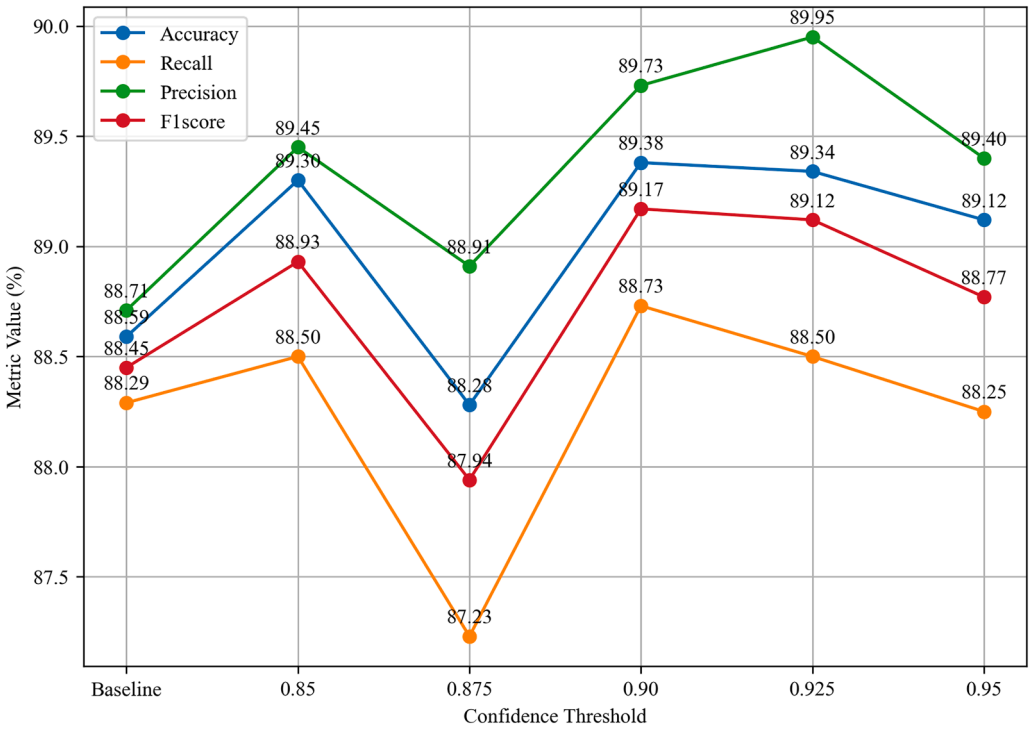

Fig. 11. Impact of five different confidence thresholds (0.85, 0.875, 0.90, 0.925, 0.95) on the performance of ResNetRS50 trained with semi-supervised learning. The Baseline represents the performance achieved by the original ResNetRS50 without the adoption of semi-supervised learning.

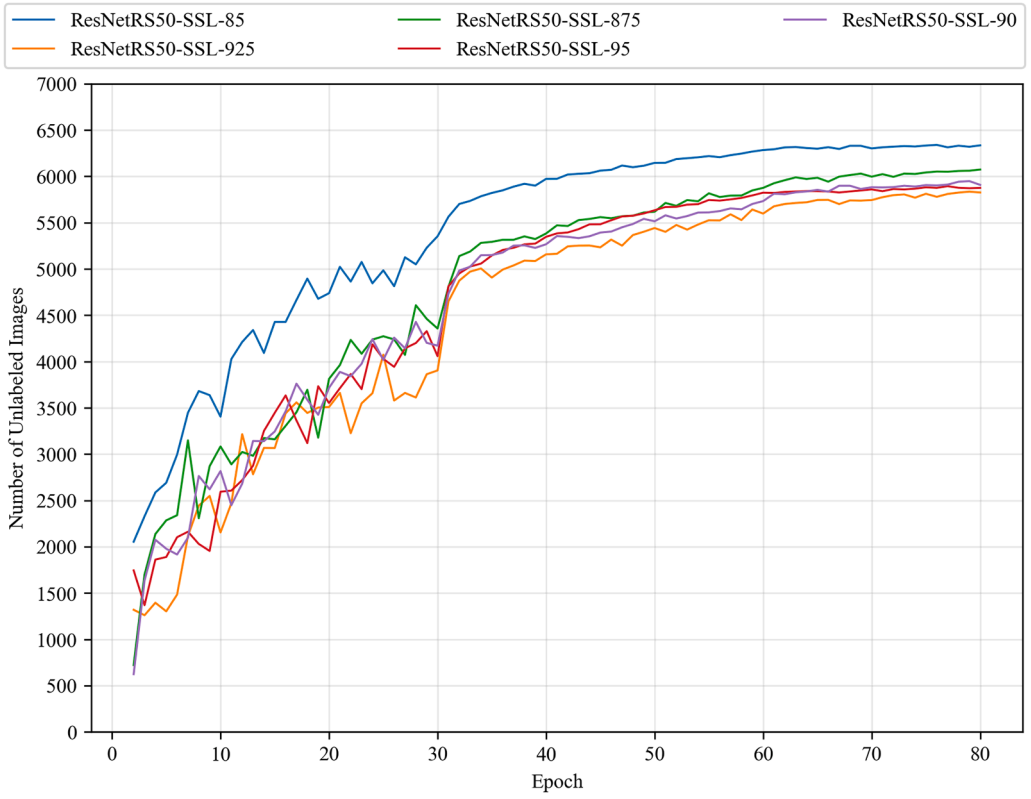

Fig. 12. Numbers of pseudo-labeled images incorporated into the training set over 80 epochs for different confidence thresholds (0.85, 0.875, 0.90, 0.925, 0.95). Each curve (ResNetRS50-SSL-85, −875, −90, −925, −95) shows how varying the threshold affects the integration of unlabeled data into training.

performed best in BBCH 3. The results suggest distinct differences in model performance depending on the specific rice growth stage.

Fig. 14 presents confusion matrices for ResNetRS50, ResNetRS50-SSL, C-ResNetRS50-SSL, and CO-ResNetRS50-SSL. In the ResNetRS50 matrix, BBCH 2 is occasionally mislabeled as BBCH 1 and BBCH 3, and a number of BBCH 3 samples appear under BBCH 4. In ResNetRS50-SSL,

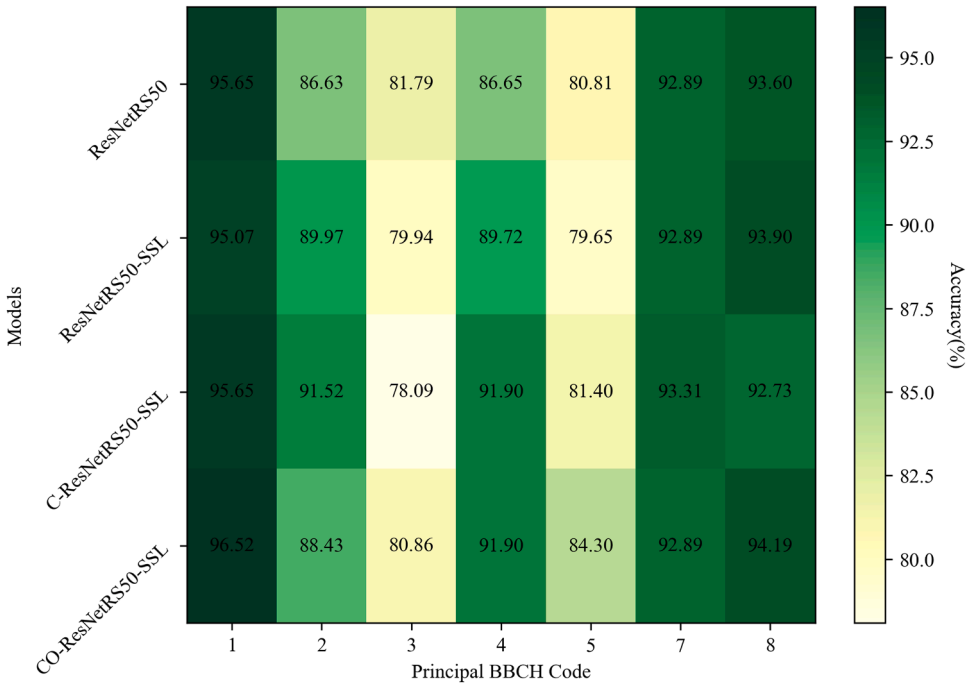

Fig. 13. Accuracy comparison of four different models (ResNetRS50, ResNetRS50-SSL, C-ResNetRS50-SSL, and CO-ResNetRS50-SSL) in identifying multiple key growth stages of rice (BBCH 1,2,3,4,5,7,8). The darker the color on the heatmap on the right, the higher the classification accuracy.

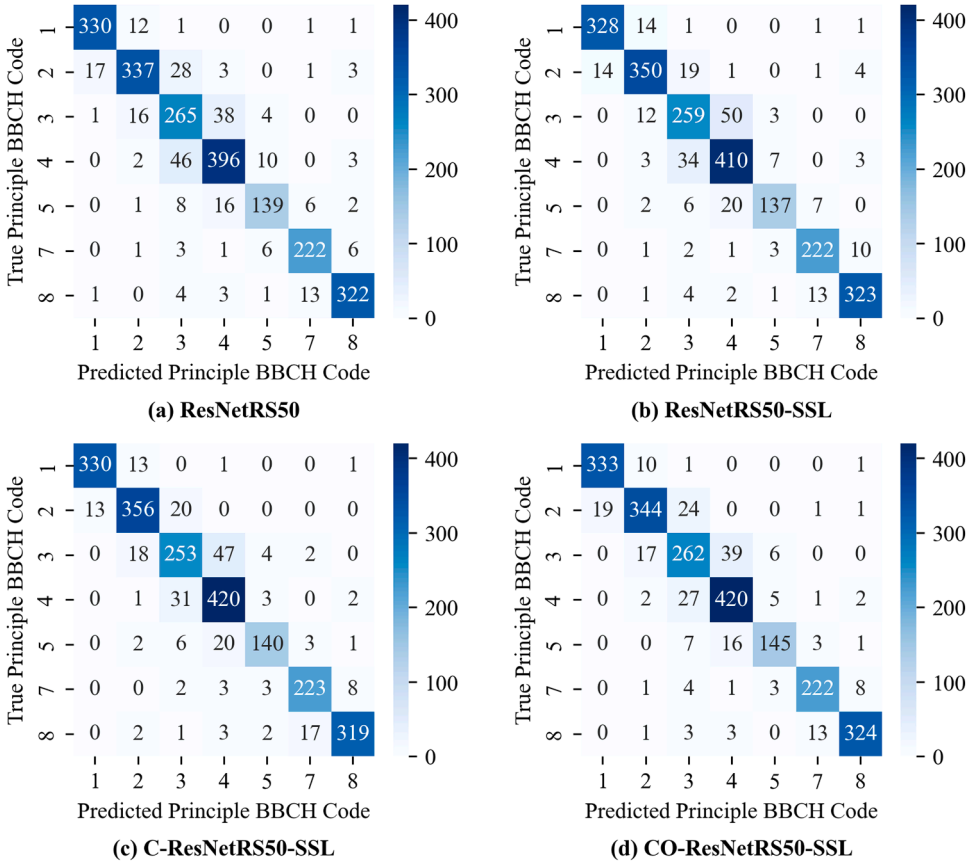

Fig. 14. Confusion matrices comparing four models—(a) ResNetRS50, (b) ResNetRS50-SSL, (c) C-ResNetRS50-SSL, and (d) CO-ResNetRS50-SSL—for rice growth stage prediction using the principle BBCH code. The vertical axis indicates the true BBCH code, and the horizontal axis shows the predicted BBCH code. Correctly classified samples lie on the main diagonal, with darker colors signifying higher classification accuracy. The color bar on the right displays the distribution of correct predictions, where greater intensity corresponds to a larger number of correctly classified samples.

similar overlaps persist among BBCH 2, 3, and 4, and certain BBCH 4 instances appear under BBCH 5. C-ResNetRS50-SSL continues to show confusion between BBCH 3 and BBCH 4, along with some BBCH 2 samples classified as BBCH 1. Meanwhile, CO-ResNetRS50-SSL also indicates misclassifications among adjacent stages, particularly those that share morphological traits (e.g., BBCH 2, 3, and 4). Overall, these patterns highlight how incremental phenotypic changes between closely related growth stages lead to classification errors across all four models.

3.5. Impacts of training image resolution on model performance

Fig. 15 summarizes the CO-ResNetRS50-SSL model's performance at three image resolutions. At 128 × 128, the model achieves 88.37 % accuracy, 85.91 % recall, 89.58 % precision, an F1 score of 87.75 %, a training duration of 1.48 hours, and an inference latency of 18.94 milliseconds. When trained at 224 × 224, it attains a significantly increased accuracy of 90.31 %, recall of 88.97 %, precision of 90.53 %, F1 score of 89.75 %, a slightly increased training duration of 2.68 hours, and latency of 19.02 milliseconds. At 512 × 512, the model records a slightly changed accuracy, recall, precision, an F1 score, a significantly increased training duration of 9.35 hours, and latency of 19.85 milliseconds. It demonstrates that higher resolutions yield only marginal gains in accuracy and recall but demand significantly more training time, whereas intermediate resolutions provide a more balanced trade-off between performance and computational cost.

4. Discussion

Accurate identification of crop growth stages is essential for optimizing management practices and scheduling agriculture activities accordingly. However, existing deep learning models face challenges, especially in complex field conditions with limited labeled data. To address these challenges, the CO-ResNetRS50-SSL model, a semi-supervised based image classification model that incorporates CA and ODConv Modules for model enhancements, was proposed, and its performance to identify the key growth stages were tested.

4.1. The benefits of semi-supervised learning

The experimental results demonstrated the clear advantage of semi-supervised learning in improving model performance. Our SSL-enhanced model, CO-ResNetRS50-SSL, consistently outperformed the baseline across all metrics, with accuracy showing the most significant improvement. These findings align with the work of Liu et al. (2023), who also observed that SSL improves model generalization by effectively leveraging both labeled and unlabeled data. Importantly, this study provides new insights into the impact of label prediction confidence on SSL's benefits. Specifically, our results show that models such as ResNetRS50-SSL-85, which incorporated a larger amount of unlabeled data with a lower CT, did not perform as well as ResNetRS50-SSL-90, which used fewer unlabeled data but with a higher CT. This indicates that the quality of pseudo-labeled data, governed by the CT, plays a more critical role in determining model accuracy than the sheer quantity of unlabeled data. This finding supports the theory that incorporating too much low-confidence data can introduce noise and reduce overall model performance, suggesting diminishing returns when the CT is too low. However, as the CT is further increased to 0.925 and 0.95, while the precision improves at a CT of 0.925, the other metrics show varying degrees of decline. This indicates that although high-confidence pseudo-labels contribute to improved precision, the reduced number of incorporated unlabeled samples limit the diversity of the training data, thereby weakening the model's generalization ability. Therefore, it can be concluded that, selecting an optimal CT is key to maximizing the effectiveness of SSL. While semi-supervised learning has demonstrated its potential to enhance model accuracy, its overall contribution in this study remains moderate. Further research is necessary to fully exploit its capabilities for achieving greater accuracy improvements.

4.2. Model enhancement with CA and ODConv modules

The ablation experiments reveal that integrating Coordinate Attention (CA) into the model results in significantly improved performance,

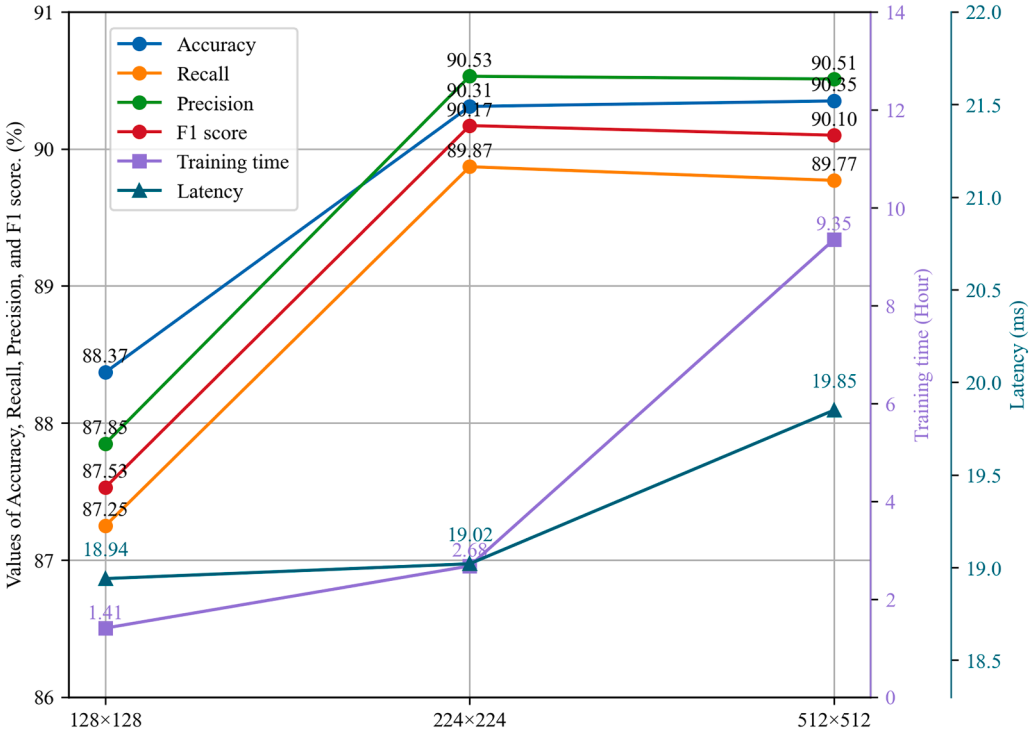

Fig. 15. Performance comparison of CO-ResNetRS50-SSL on the datasets with three different training image resolutions (128 × 128, 224 × 224, and 512 × 512), evaluated using metrics including accuracy, recall, precision, F1 score, training time, and latency.

particularly in its ability to identify and classify features in complex datasets. The CA module enhances the network's capacity to capture and interpret the relative positional and structural relationships between objects in an image. This improvement is crucial for classification tasks where backgrounds are complex (Zhao et al., 2024a), such as in crop fields, where distinguishing crops from their environment is essential (Pochet et al., 2023). By encoding positional information effectively, CA not only strengthens the model's ability to capture long-range dependencies but also mitigates the limitations observed in other attention mechanisms like BAM (Park et al., 2018) and CBAM (Woo et al., 2018), which are less effective in modeling these relationships. Studies have shown that models equipped with positional encoding, such as CA, can significantly enhance the interpretability and representational power of neural networks in spatially complex tasks (Hou et al., 2021).

Additionally, the introduction of Omni-Dimensional Dynamic Convolution (ODConv) further boosts the model's performance by allowing the convolutional kernels to dynamically adapt to varying feature sizes, shapes, and orientations within the data (Li et al., 2022). Traditional convolutional layers often struggle to accommodate this variability, particularly in scenarios where the target features exhibit significant diversity, such as in agricultural fields where crops may appear at different scales and conditions. ODConv enhances the flexibility of the convolutional operation by dynamically adjusting the convolutional filters to better fit the target features, thereby increasing the network's capacity to generalize across different instances. This adaptability is especially advantageous in different field conditions, where variations in crop appearance are influenced by growth stages, environmental factors, and camera perspectives. Nevertheless, ODConv also increases the model's parameter count. Unlike traditional convolutions, which use fixed convolutional kernels, ODConv requires additional parameters to store the dynamic adjustment mechanisms, which may reduce the model's real-time performance and practicality, especially when computational resources are limited.

From a theoretical perspective, the improvements seen with CA and ODConv can be explained by their complementary roles in feature extraction. CA enhances the model's sensitivity to spatial structures by embedding positional cues directly into the feature maps, thereby guiding the model in distinguishing between objects in the image more effectively (Hou et al., 2021). ODConv, on the other hand, increases the model's flexibility by allowing the convolutional kernels to better align with the inherent variability in the data, which is crucial when handling complex and dynamic agricultural scenes. Together, these two modules address the limitations of traditional convolutions, which often assume fixed, context-independent kernel operations, and attention mechanisms, which may fail to model positional dependencies effectively. While CA and ODConv demonstrated exceptional performance in various applications, their efficacy may be constrained in scenarios where the distinction between target features and features of other classes is minimal.

4.3. The model performance across key growth stages

Despite attaining superior overall accuracy, CO-ResNetRS50-SSL exhibited certain weaknesses, particularly in distinguishing the stem elongation (BBCH 3) and heading (BBCH 5) stages. These stages are characterized by subtle morphological shifts that can overlap with adjacent growth phases, such as booting (BBCH 4), making classification difficult (Qin et al., 2023). In the case of heading (BBCH 5), the limited sample size further complicates classification. The model is more prone to bias toward stages with greater representation in the dataset, leading to errors in minority class recognition, as noted by Bailly et al., (2022). Another challenge arises from the visual prominence of the panicle, which emerges in the heading stage. This feature, critical for distinguishing heading from booting, may be obscured in images taken at certain angles or lighting conditions. In agronomic terms, the proper identification of heading is crucial as it marks the transition to reproductive development, influencing grain yield potential and management interventions.

Compared to the apple datasets in Liu et al.(2023), the indistinct morphological changes between rice growth stages—especially those involving dynamic vegetative to reproductive transitions—make rice classification more complex. The nuanced features in rice, such as tiller angle and panicle emergence, require more sophisticated feature extraction methods for accurate growth stage identification. While Tan et al. (2023) achieved high accuracy by focusing on just three stages (booting, heading, and filling), their reduced scope simplifies the task compared to this study, which covers a broader range of stages with more complex variations. The present study highlights the need for enhanced feature extraction and more balanced datasets to improve accuracy in complex growth stages critical for agronomic decision-making. Our findings suggest that improving classification accuracy for complex stages like BBCH 3 and BBCH 5 may involve addressing sample imbalance, enhancing feature extraction for subtle morphological changes, and expanding the dataset for underrepresented stages. Agronomically, this would allow for more precise monitoring and decision-making during critical growth periods, leading to better crop management and yield optimization. Additionally, while the proposed method outperforms others across various growth stages, its accuracy in identifying BBCH 3 and BBCH 5 remains relatively lower compared to other stages, indicating the need for further improvements.

Overall, while CO-ResNetRS50-SSL provides strong performance across most stages, challenges remain in distinguishing closely related growth phases such as BBCH 3 and BBCH 5. Future refinements—like improving sample balance, augmenting features for subtle morphological changes, and increasing dataset diversity—could further enhance classification accuracy in these complex scenarios. In practical terms, more reliable predictions of critical stages would allow farmers and agronomists to optimize their interventions, ultimately leading to improved crop management and higher yields.

4.4. Limitations

Despite the superiority achieved by our method, there are still two limitations to address in the future work. First, in our SSL based approach, we applied a CT to filter out low-confidence pseudo-labeled images (PLIMs), which might not be sufficient to filter out noisy labels and ensure a balanced distribution of PLIMs across different growth stages. Therefore, it could struggle to completely mitigate the impact of noisy labels. This limitation may exacerbate the issue of BBCH 5 having fewer samples than other growth stages, reducing the model's ability to generalize to under-represented categories. Second, to balance computational efficiency, we selected a resolution of 224×224. While this resolution effectively reduced the computational cost, it may have resulted in the loss of some fine details compared to the higher resolution of 512×512, potentially affecting the model's ability to capture subtle features.

To address these issues, future research could explore refining both the pseudo-labeling strategy and the image resolution methodology. First, we will explore advanced confidence thresholding techniques—such as adaptive thresholding and uncertainty estimation—to better filter out low-confidence pseudo-labeled images and mitigate the impact of noisy labels. It could ensure a more balanced distribution of pseudo-labeled samples across all growth stages, particularly addressing the under-representation of BBCH 5. Second, we intend to investigate multi-scale or efficient high-resolution processing strategies that allow the use of higher resolution inputs (e.g., 512×512) without significantly increasing computational costs. Such strategies may enable the model to capture finer details and subtle features, ultimately enhancing its generalization capability under complex field conditions.

5. Conclusions

This study proposes CO-ResNetRS50-SSL, a semi-supervised learning model based on the improved ResNetRS architecture for rice growth stage classification. By incorporating SSL for utilizing unlabeled data, the model improves generalization ability and reduces dependency on large scale labeled data. Meanwhile, the integration of the CA and ODConv Modules enhances the model's feature extraction and target perception capabilities. Experimental results show that the CO-ResNetRS50-SSL model achieves a classification accuracy of 90.38 %, surpassing other mainstream models. Due to the improved accuracy and generalization as well as the reduced need for manual labeling by utilizing unlabeled data, this method is particularly suitable for data-scarce and cost-sensitive fields, such as agriculture. In rice field monitoring, the SSL-based automated system can reduce time and financial costs, allowing for rapid assessment of crop growth stages and optimization of management practices. Furthermore, integrating this model into smart agricultural robots allows for automated crop management, enhancing operational efficiency and practical applicability. While significant results have been achieved, future work will focus on exploring adaptive thresholding to better filter pseudo-labeled images and investigate efficient multi-scale high-resolution processing methods to capture fine details while maintaining computational efficiency.

CRediT authorship contribution statement

Liang Zeyun: Writing – review & editing, Visualization, Validation, Software. **Yang Guangpeng:** Writing – original draft, Software, Conceptualization. **Yan Changqing:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Methodology, Conceptualization. **Zhao Gang:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization. **Yu Qiang:** Writing – review & editing, Conceptualization. **Srivastava Amit Kumar:** Writing – review & editing, Validation. **Wu Genghong:** Writing – review & editing, Validation. **Cheng Han:** Writing – review & editing, Validation, Software.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is supported by the Shaanxi Key R&D Program Project (grant no. 2023-ZDLNY-64) and the MOE (Ministry of Education in China) Liberal Arts and Social Sciences Foundation of China (Grant No. 23YJC790122).

Data availability

Data will be made available on request.

References

Aich, S., Josuttes, A., Ovsyannikov, I., Strueby, K., Ahmed, I., Duddu, H.S., Pozniak, C., Shirtliffe, S., Stavness, I., 2018. Deepwheat: Estimating Phenotypic Traits from Crop Images with Deep Learning. Presented at the 2018 IEEE Winter conference on applications of computer vision (WACV), IEEE, pp. 323–332. ⟨https://doi.org/10.1109/WACV.2018.00042⟩.

Alabsi, B., Anbar, M., Rihan, S., 2023. CNN-CNN: Dual Convolutional Neural Network Approach for Feature Selection and Attack Detection on Internet of Things Networks. Sensors 23, 6507. https://doi.org/10.3390/s23146507.

Amorim, W.P., Tetila, E.C., Pistori, H., Papa, J.P., 2019. Semi-Supervised Learning with Convolutional Neural Networks for UAV Images Automatic Recognition. Comput. Electron. Agric. 164, 104932. https://doi.org/10.1016/j.compag.2019.104932.

Bailly, A., Blanc, C., Francis, É ., Guillotin, T., Jamal, F., Wakim, B., Roy, P., 2022. Effects of Dataset Size and Interactions on the Prediction Performance of Logistic Regression and Deep Learning Models. Comput. Methods Prog. Biomed. 213, 106504. https://doi.org/10.1016/j.cmpb.2021.106504.

Bello, I., Fedus, W., Du, X., Cubuk, E.D., Srinivas, A., Lin, T.Y., Shlens, J., Zoph, B., 2021. Revisiting ResNets: Improved Training and Scaling Strategies. arXiv E-Prints arXiv 2103, 07579. https://doi.org/10.48550/arXiv.2103.07579.

Benchallal, F., Hafiane, A., Ragot, N., Canals, R., 2024. ConvNeXt Based Semi-Supervised Approach with Consistency Regularization for Weeds Classification. Expert Syst. Appl. 239, 122222. https://doi.org/10.1016/j.eswa.2023.122222.

de Castro Pereira, R., Hirose, E., Ferreira de Carvalho, O.L., da Costa, R.M., Borges, D.L., 2022. Detection and Classification of Whiteflies and Development Stages on Soybean Leaves Images Using an Improved Deep Learning Strategy. Comput. Electron. Agric. 199, 107132. https://doi.org/10.1016/j.compag.2022.107132.

Chen, J., Kao, S., He, H., Zhuo, W., Wen, S., Lee, C.H., Chan, S.H.G., 2023. Run, Don't Walk: Chasing Higher FLOPS for Faster Neural Networks, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12021–12031. ⟨https://doi.org/10.1109/CVPR52729.2023.01157⟩.

Coleman, G.R.Y., Kutugata, M., Walsh, M.J., Bagavathiannan, M.V., 2024. Multi-growth Stage Plant Recognition: A Case Study of Palmer Amaranth (*Amaranthus palmeri*) in Cotton (*Gossypium hirsutum*). Comput. Electron. Agric. 217, 108622. https://doi.org/10.1016/j.compag.2024.108622.

Cortinas, E., Emmi, L., Gonzalez-de-Santos, P., 2023. Crop Identification and Growth Stage Determination for Autonomous Navigation of Agricultural Robots. Agronomy 13, 2873. https://doi.org/10.3390/agronomy13122873.

Deng, P., Jiang, Z., Ma, H., Rao, Y., Zhang, W., 2025. Improving Crop Image Recognition Performance Using Pseudolabels. Inf. Process. Agric. 12, 17–26. https://doi.org/10.1016/j.inpa.2024.02.001.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2020. An Image is Worth 16× 16 Words: Transformers for Image Recognition at Scale. arXiv 2010, 11929. https://doi.org/10.48550/arXiv.2010.11929.

Ferentinos, K.P., 2018. Deep Learning Models for Plant Disease Detection and Diagnosis. Comput. Electron. Agric. 145, 311–318. https://doi.org/10.1016/j.compag.2018.01.009.

Ge, H., Ma, F., Li, Z., Tan, Z., Du, C., 2021. Improved Accuracy of Phenological Detection in Rice Breeding by Using Ensemble Models of Machine Learning Based on UAV-RGB Imagery. Remote Sens 13, 2678. https://doi.org/10.3390/rs13142678.

Guo, J., Jia, N., Bai, J., 2022. Transformer Based on Channel-spatial Attention for Accurate Classification of Scenes in Remote Sensing Image. Sci. Rep. 12, 15473. https://doi.org/10.1038/s41598-022-19831-z.

Guo, J., Yang, Y., Lin, X., Memon, M.S., Liu, W., Zhang, M., Sun, E., 2023. Revolutionizing Agriculture: Real-Time Ripe Tomato Detection with the Enhanced Tomato-YOLOv7 System. IEEE Access 11, 133086–133098. https://doi.org/10.1109/ACCESS.2023.3336562.

Hou, Q., Zhou, D., Feng, J., 2021. Coordinate Attention for Efficient Mobile Network Design, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Nashville, TN, USA, pp. 13708–13717. https://doi.org/10.1109/CVPR46437.2021.01350.

Hu, Y., Meng, A., Wu, Y., Zou, L., Jin, Z., Xu, T., 2023. Deep-agriNet: A Lightweight Attention-based Encoder-decoder Framework for Crop Identification Using Multispectral Images. Front. Plant Sci. 14, 1124939. https://doi.org/10.3389/fpls.2023.1124939.

Janiesch, C., Zschech, P., Heinrich, K., 2021. Machine Learning and Deep Learning. Electron. Mark. 31, 685–695. https://doi.org/10.1007/s12525-021-00475-2.

Jia, Y., Fu, K., Lan, H., Wang, X., Su, Z., 2024. Maize Tassel Detection with CA-YOLO for UAV Images in Complex Field Environments. Comput. Electron. Agric. 217, 108562. https://doi.org/10.1016/j.compag.2023.108562.

Khan, S., Tufail, M., Khan, M.T., Khan, Z.A., Iqbal, J., Alam, M., 2021. A Novel Semi-Supervised Framework for UAV Based Crop/Weed Classification. PLOS ONE 16, e0251008. https://doi.org/10.1371/journal.pone.0251008.

Lancashire, P.D., Bleiholder, H., Boom, T., van den, Langelüddeke, P., Stauß, R., Weber, E., Witzenberger, A., 1991. A Uniform Decimal Code for Growth Stages of Crops and Weeds. Ann. Appl. Biol. 119, 561–601. https://doi.org/10.1111/j.1744-7348.1991.tb04895.x.

Li, C., Zhou, A., Yao, A., 2022. Omni-Dimensional Dynamic Convolution. arXiv 2209, 07947. https://doi.org/10.48550/arXiv.2209.07947.

Li, Y., Chao, X., 2021. Semi-supervised Few-shot Learning approach for Plant Diseases Recognition. Plant Methods 17, 68. https://doi.org/10.1186/s13007-021-00770-1.

Liu, H., Xu, Z., 2023. Editorial: Machine Vision and Machine Learning for Plant Phenotyping and Precision Agriculture. Front. Plant Sci. 14, 1331918. https://doi.org/10.3389/fpls.2023.1331918.

Liu, Y., Gao, W., He, P., Tang, J., Hu, L., 2023. Apple Phenological Period Identification in Natural Environment Based on Improved ResNet50 Model. Smart Agric. 5, 13. https://doi.org/10.12133/j.smartag.SA202304009.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Presented at the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Montreal, QC, Canada, pp. 9992–10002. ⟨https://doi.org/10.1109/ICCV48922.2021.00986⟩.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A ConvNet for the 2020 s, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2022 IEEE/CVF Conference on Computer Vision and

Pattern Recognition (CVPR), IEEE, New Orleans, LA, USA, pp. 11966–11976. ⟨https://doi.org/10.1109/CVPR52688.2022.01167⟩.
Lyu, M., Lu, X., Shen, Y., Tan, Y., Wan, L., Shu, Q., He, Yuhong, He, Yong, Cen, H., 2023. UAV Time-series Imagery with Novel Machine Learning to Estimate Heading Dates of Rice Accessions for Breeding. Agric. For. Meteor. 341, 109646. https://doi.org/10.1016/j.agrformet.2023.109646.
Ma, N., Zhang, X., Zheng, H.T., Sun, J., 2018. ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.), Computer Vision – ECCV 2018, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 122–138. https://doi.org/10.1007/978-3-030-01264-9_8.
Mao, A., Mohri, M., Zhong, Y., 2023. Cross-Entropy Loss Functions: Theoretical Analysis and Applications, in: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (Eds.), Proceedings of the 40th International Conference on Machine Learning, Proceedings of Machine Learning Research. PMLR, pp. 23803–23828. ⟨https://proceedings.mlr.press/v202/mao23b.html⟩.
Nyéki, A., Neményi, M., 2022. Crop Yield Prediction in Precision Agriculture. Agronomy 12, 2460. https://doi.org/10.3390/agronomy12102460.
Panda, M.K., Sharma, A., Bajpai, V., Subudhi, B.N., Thangaraj, V., Jakhetiya, V., 2022. Encoder and Decoder Network with ResNet-50 and Global Average Feature Pooling for Local Change Detection. Comput. Vis. Image Under 222, 103501. https://doi.org/10.1016/j.cviu.2022.103501.
Park, J., Woo, S., Lee, J.Y., Kweon, I.S., 2018. BAM: Bottleneck Attention Module. arXiv 1807, 06514. https://doi.org/10.48550/arXiv.1807.06514.
Pochet, E., Maroun, R., Trullo, R., 2023. RoFormer for Position Aware Multiple Instance Learning in Whole Slide Image Classification. arXiv Prepr. arXiv 2310, 01924. https://doi.org/10.48550/arXiv.2310.01924.
Qin, J., Hu, T., Yuan, J., Liu, Q., Wang, W., Liu, J., Guo, L., Song, G., 2023. Deep-Learning-Based Rice Phenological Stage Recognition. Remote Sens 15, 2891. https://doi.org/10.3390/rs15112891.
Ravlić, M., Kulundžić, A.M., Baličević, R., Marković, M., Vuletić, M.V., Kranjac, D., Sarajlić, A., 2022. Allelopathic Potential of Sunflower Genotypes at Different Growth Stages on Lettuce. Appl. Sci. 12, 12568. https://doi.org/10.3390/app122412568.
Roy, A.M., Bhaduri, J., 2022. Real-time Growth Stage Detection Model for High Degree of Occultation Using DenseNet-fused YOLOv4. Comput. Electron. Agric. 193, 106694. https://doi.org/10.1016/j.compag.2022.106694.
Schieck, M., Krajsic, P., Loos, F., Hussein, A., Franczyk, B., Kozierkiewicz, A., Pietranik, M., 2023. Comparison of Deep Learning Methods for Grapevine Growth Stage Recognition. Comput. Electron. Agric. 211, 107944. https://doi.org/10.1016/j.compag.2023.107944.
Sheng, R.T.C., Huang, Y.H., Chan, P.C., Bhat, S.A., Wu, Y.C., Huang, N.F., 2022. Rice Growth Stage Classification via RF-Based Machine Learning and Image Processing. Agriculture 12, 2137. https://doi.org/10.3390/agriculture12122137.
Sun, M., Dai, Y., Zhang, S., Liang, H., 2025. Risk Assessment of Extreme Drought and Extreme Wetness During Growth Stages of Major Crops in China. Sustainability 17, 2221. https://doi.org/10.3390/su17052221.
Tan, S., Lu, H., Yu, J., Lan, M., Hu, X., Zheng, H., Peng, Y., Wang, Y., Li, Z., Qi, L., Ma, X., 2023. In-field Rice Panicles Detection and Growth Stages Recognition Based on RiceRes2Net. Comput. Electron. Agric. 206, 107704. https://doi.org/10.1016/j.compag.2023.107704.
Tian, Q., Zhao, G., Yan, C., Yao, L., Qu, J., Yin, L., Feng, H., Yao, N., Yu, Q., 2024. Enhancing Practicality of Deep Learning for Crop Disease Identification Under Field Conditions: Insights from Model Evaluation and Crop-specific Approaches. Pest Manag. Sci. ps. 8317. https://doi.org/10.1002/ps.8317.
Wang, B., Liu, Y., Sheng, Q., Li, J., Tao, J., Yan, Z., 2022. Rice Phenology Retrieval Based on Growth Curve Simulation and Multi-Temporal Sentinel-1 Data. Sustainability 14, 8009. https://doi.org/10.3390/su14138009.
Wang, D., Wang, X., Chen, Y., Wu, Y., Zhang, X., 2023. Strawberry Ripeness Classification Method in Facility Environment Based on Red Color Ratio of Fruit Rind. Comput. Electron. Agric. 214, 108313. https://doi.org/10.1016/j.compag.2023.108313.
Woo, S., Park, J., Lee, J.Y., Kweon, I.S., 2018. CBAM: Convolutional Block Attention Module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.), Computer Vision – ECCV 2018. Springer International Publishing, Cham, pp. 3–19. https://doi.org/10.1007/978-3-030-01234-2_1.
Xiao, K., Zhou, L., Yang, H., Yang, L., 2022. Phalaenopsis Growth Phase Classification Using Convolutional Neural Network. Smart Agric. Technol. 2, 100060. https://doi.org/10.1016/j.atech.2022.100060.
Yan, C., Liang, Z., Yin, L., Wei, S., Tian, Q., Li, Y., Cheng, H., Liu, J., Yu, Q., Zhao, G., Qu, J., 2024. AFM-YOLOv8s: An Accurate, Fast and Highly Robust Model for Detection of Sporangia of Plasmopara Viticola with Various Morphological Variants. Plant Phenomics 6, 0246. https://doi.org/10.34133/plantphenomics.0246.
Yu, F., Zhang, Q., Xiao, J., Ma, Y., Wang, M., Luan, R., Liu, X., Ping, Y., Nie, Y., Tao, Z., Zhang, H., 2023. Progress in the Application of CNN-Based Image Classification and Recognition in Whole Crop Growth Cycles. Remote Sens 15, 2988. https://doi.org/10.3390/rs15122988.
Yue, Y., Li, J.H., Fan, L.F., Zhang, L.L., Zhao, P.F., Zhou, Q., Wang, N., Wang, Z.Y., Huang, L., Dong, X.H., 2020. Prediction of Maize Growth Stages Based on Deep Learning. Comput. Electron. Agric. 172, 105351. https://doi.org/10.1016/j.compag.2020.105351.
Zhang, Y., Xiao, D., Liu, Y., 2021. Automatic Identification Algorithm of the Rice Tiller Period Based on PCA and SVM. IEEE Access 9, 86843–86854. https://doi.org/10.1109/ACCESS.2021.3089670.
Zhao, B., Chen, Y., Jia, X., Ma, T., 2024a. Steel Surface Defect Detection Algorithm in Complex Background Scenarios. Measurement 237, 115189. https://doi.org/10.1016/j.measurement.2024.115189.
Zhao, G., Zhao, Q., Webber, H., Johnen, A., Rossi, V., Junior, A.F.N., 2024b. Integrating Machine Learning and Change Detection for Enhanced Crop Disease Forecasting in Rice Farming: A Multi-regional Study. Eur. J. Agron. 160, 127317. https://doi.org/10.1016/j.eja.2024.127317.