Plant Methods

METHODOLOGY

Open Access

A deep learning approach for deriving wheat phenology from near-surface RGB image series using spatiotemporal fusion

Yucheng Cai[1,2†], Yan Li[1,2†], Xuerui Qi[1,2], Jianqing Zhao[3], Li Jiang[4], Yongchao Tian[1,5], Yan Zhu[1,2], Weixing Cao[1,2] and Xiaohu Zhang[1,2,5*]

Abstract

Accurate monitoring of wheat phenological stages is essential for effective crop management and informed agricultural decision-making. Traditional methods often rely on labour-intensive field surveys, which are prone to subjective bias and limited temporal resolution. To address these challenges, this study explores the potential of near-surface cameras combined with an advanced deep-learning approach to derive wheat phenological stages from high-quality, real-time RGB image series. Three deep learning models based on three different spatiotemporal feature fusion methods, namely sequential fusion, synchronous fusion, and parallel fusion, were constructed and evaluated for deriving wheat phenological stages with these near-surface RGB image series. Moreover, the impact of different image resolutions, capture perspectives, and model training strategies on the performance of deep learning models was also investigated. The results indicate that the model using the sequential fusion method is optimal, with an overall accuracy (OA) of 0.935, a mean absolute error (MAE) of 0.069, F1-score (F1) of 0.936, and kappa coefficients (Kappa) of 0.924 in wheat phenological stages. Besides, the enhanced image resolution of 512 × 512 pixels and a suitable image capture perspective, specifically a sensor viewing angle of 40° to 60° vertically, introduce more effective features for phenological stage detection, thereby enhancing the model's accuracy. Furthermore, concerning the model training, applying a two-step fine-tuning strategy will also enhance the model's robustness to random variations in perspective. This research introduces an innovative approach for real-time phenological stage detection and provides a solid foundation for precision agriculture. By accurately deriving critical phenological stages, the methodology developed in this study supports the optimization of crop management practices, which may result in improved resource efficiency and sustainability across diverse agricultural settings. The implications of this work extend beyond wheat, offering a scalable solution that can be adapted to monitor other crops, thereby contributing to more efficient and sustainable agricultural systems.

Keywords Wheat, Phenology monitoring, RGB image series, Deep learning, Spatiotemporal feature

[†]Yucheng Cai and Yan Li have authors contributed equally to this work.

*Correspondence:
Xiaohu Zhang
zhangxiaohu@njau.edu.cn
Full list of author information is available at the end of the article

Introduction

Wheat is a widely cultivated and consumed cereal crop in the world, and the management practices for its cultivation rely on monitoring its phenological stages [19]. Accurate monitoring of wheat phenological stages can optimize field management, predict yield and harvest times, facilitate pest and disease control, and adjust planting structures, thereby playing a crucial role in enhancing planting efficiency [32, 33]. The traditional monitoring of wheat phenological stages often relies on manual field surveys, which consume a significant amount of labor and suffer from subjective biases. The development of crop growth monitoring platforms and intelligent algorithms has led to the investigation of various methods for obtaining wheat phenological stages information, including satellite [22] and unmanned aerial vehicles [44]. Although satellite remote sensing platforms can acquire images covering extensive areas, their temporal resolution is low, making it challenging to obtain high-temporal-resolution image series [16]. The use of unmanned aerial vehicle platforms for data collection is limited by the inherent difficulties of acquiring images in bad weather conditions [7]. Conversely, near-surface platforms equipped with cameras offer a practical solution by continuously capturing high-resolution image series throughout the day and under all weather conditions [38, 55]. Moreover, their low-cost, convenient operation renders them an invaluable tool for monitoring wheat phenological stages [21, 24, 54].

In addition, recent advances in deep learning have led to remarkable progress in the field of agriculture [17, 29] offering new solutions for complex tasks such as wheat spike detection [57], pest and disease detection [42] and yield prediction [50]. The advancement can be attributed to the efforts of researchers who have been actively engaged in the collection and construction of new datasets, as well as the examination of the characteristics of these data [40]. Furthermore, the development of novel model architectures built on agricultural datasets has been instrumental in this progression [56]. However, research on crop phenological stage detection using deep learning remains very limited. Previous studies frequently concentrated on identifying specific phenological stages and used single-stage images as input for detection models [25, 43, 55]. However, single-stage images are unable to fully capture the changes in crop phenology characteristics throughout the phenological stages, resulting in low classification accuracy[48, 51]. Although some studies incorporate temporal features into deep learning models, they fail to consider the relationship between these features and spatial features [38, 55]. Consequently, there is currently no effective wheat phenological stages detection model that can seamlessly integrate spatial features and temporal features to achieve real-time detection of wheat phenological stages.

To address the limitations of existing methods for crop phenology detection, our study employed near-surface cameras to collect a comprehensive dataset of wheat phenological stages throughout the growth period. By introducing advanced spatiotemporal feature fusion techniques, we constructed and optimized a detection model that overcomes the shortcomings of single-stage image analysis. This approach markedly improves the accuracy of phenological stage detection, facilitating monitoring and enhanced generalization across diverse conditions. Consequently, it offers a robust and efficient solution for precise crop monitoring.

Materials and methods

Data collection and preprocessing

Study area and near-surface camera image acquisition

The study was conducted from November 29, 2022, to June 3, 2023, at Baima Experimental Station of Nanjing Agricultural University in Lishui District, Nanjing City, Jiangsu Province, China (119°09′ E, 31°37′ N). The experiment was conducted using six plots, each measuring 30 m in length and 3 m in width. For data collection, a RGB camera (Hikvision E DS-2DE4223IW-D/GLT/XM, Hangzhou Hikvision Digital Technology Co., Ltd., China) was employed to capture wheat images with a resolution of 1920×1080 pixels from 8:00 to 17:00 daily, at a height of 3 m above the ground. The spatial resolution of the images was 0.05 cm per pixel, with a vertical viewing angle ranging from 20° to 80°, which was varied manually according to the plot. Data collection spanned 107 days over the entire growing season, during which 450 images were taken from each plot daily. Images collected at different time stamps within a day were labelled as the same phenological stage. Concurrently, the study recorded the commencement dates of all phenological stages, from emergence to maturity, in accordance with the established definition of wheat phenological stages [58].

Image datasets preprocessing

Since the input to the model consists of image series, this study preprocessed the images to construct a standardized image series dataset. The original images were cropped to standard images of 1000×1000 pixels. Each image was then manually annotated with phenological stage labels, and a set of time series of t image samples was created to describe the dynamic characteristics of wheat phenological stages (Fig. 1). Constructing image time series samples involved three steps.

Step 1: Randomly select an image i and place it at the t-th position in the time series, using the growth

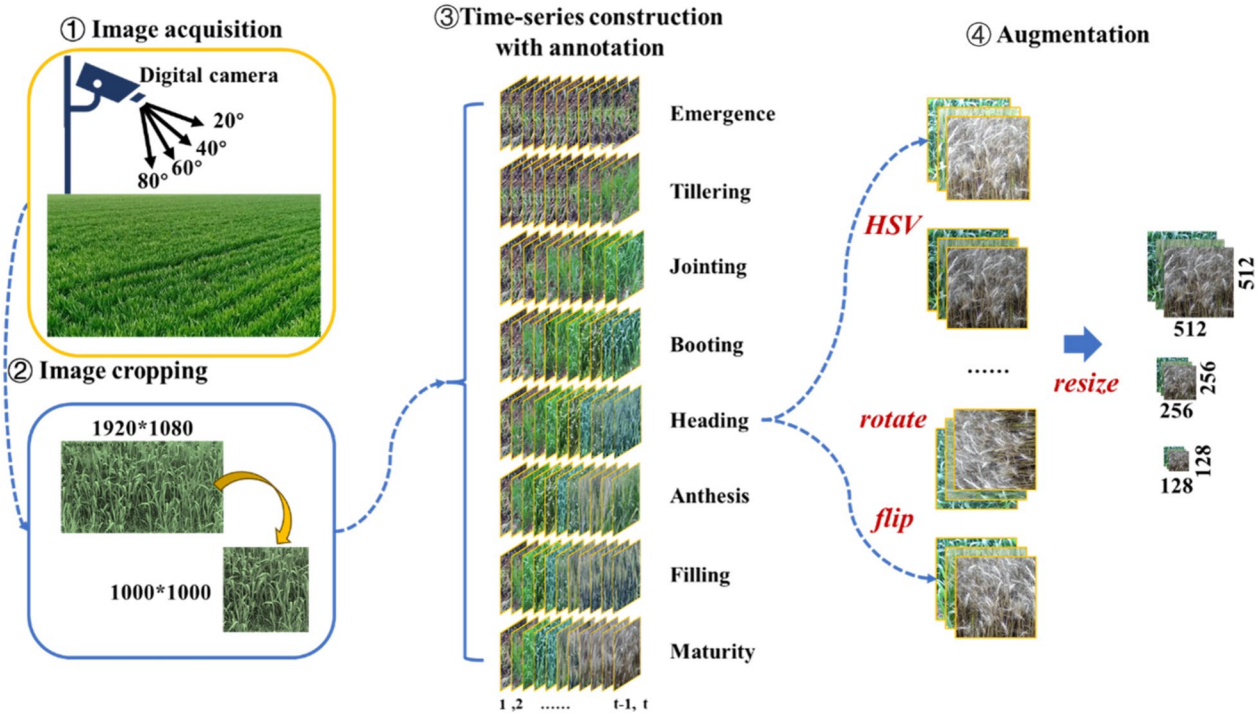

Fig.1 The data collection and preprocessing diagram

stage label *l* of the selected image as the label for this time series sample.

Step 2: Fill the positions *2/3t+1* to *t-1* in the time series with sequential images preceding the timestamp of image *i* and assign the same labels as image *i* to these images.

Step 3: Fill the remaining positions *1* to *2/3t* in the time series with sequential images prior to phenological stage *l*, where each image is from different phenological stages from the first stage to stage *l-1*. In other words, for a time series sample containing *t* images, with the phenological stage label *l*, these images represent different phenological stages from the first stage to stage *l-1*, arranged in chronological order.

In this study, *t* was set to 30. The image series samples were split into training, validation, and test sets in a 6:2:2 ratio. Furthermore, data augmentation techniques, including random rotation, flipping, and brightness adjustment, were applied to the training set (Fig. 1), resulting in a total of 13,648 image series samples (Table 1). Furthermore, in order to investigate the impact of image resolution on the model, three different resolution datasets were constructed. These comprised a low-resolution dataset with 128×128 pixels, a

Table 1 The number of image series of different wheat phenological stages

Phenology stage	Train set (Before augmentation)	Train set (After augmentation)	Validate set	Test set	Overall (After augmentation)
Emergence	612	1836	204	204	2244
Tillering	716	2148	254	254	2656
Jointing	490	1470	163	163	1796
Booting	388	1164	112	112	1388
Heading	370	1110	121	121	1352
Anthesis	284	852	98	98	1048
Filling	366	1098	105	105	1308
Maturity	496	1488	184	184	1856

medium-resolution dataset with 256×256 pixels, and a high-resolution dataset with 512×512 pixels.

Methods

In this study, the Residual network (ResNet) was selected as the baseline network [14]. Three different spatiotemporal feature fusion methods, namely sequential fusion (2.2.1), synchronous fusion (2.2.2), and parallel fusion (2.2.3), were integrated to construct three different detection models (Fig. 2). All three deep learning models incorporated the self-attention mechanism non-local module [41], which is a non-local attention model that weights and sums the features across the entire input space to capture global information. Each fusion method employs two training strategies: training from scratch and fine-tuning. Training from scratch involves initializing the model parameters randomly and training the model from the beginning using the training dataset. In contrast, fine-tuning refers to the process of taking a pre-trained model, one that has been previously trained on a larger dataset, and further training it on a smaller, task-specific dataset. Fine-tuning typically involves modifying

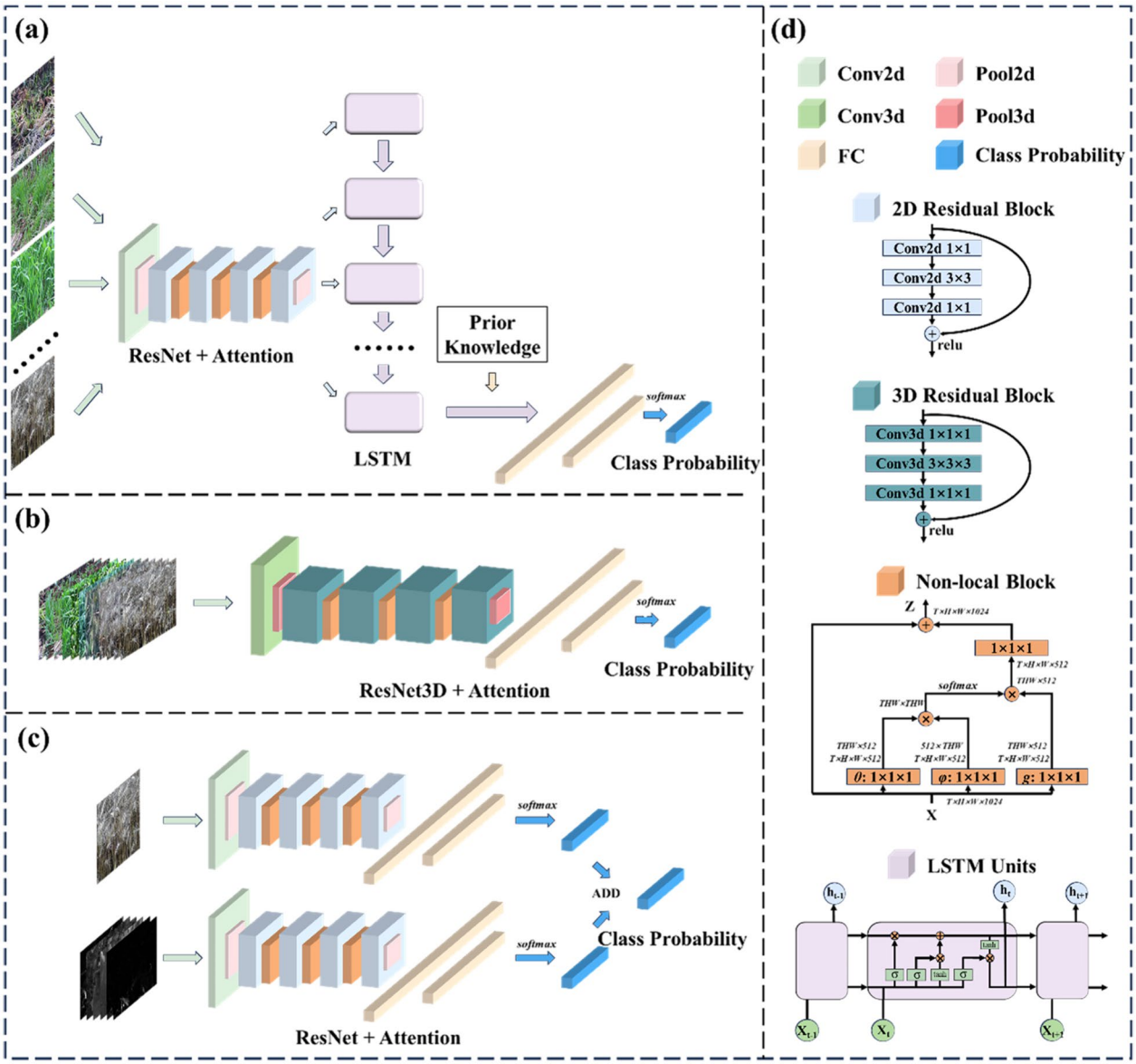

Fig.2 The architecture of three fusion methods. **a** Sequential Fusion. **b** Synchronous Fusion. **c** Parallel Fusion. **d** The description of each symbol used to represent the network architecture in (**a**) (**b**) (**c**). The prior knowledge concerns the temporal sequence of phenological stage labels. With the exception of the initial stage, the final time node's phenological stage label is always subsequent in time to that of the intermediate time node

the learning rates and optimization parameters specific to the new task, and it can be done by either retraining the entire network or updating only specific layers [15, 39].

Sequential fusion

Sequential fusion employs a model architecture that combines Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) in order to achieve the fusion of spatiotemporal features. The CNN is responsible for feature extraction, while the LSTM is tasked with feature memorization [4]. In this study, ResNet50 was selected as the convolutional neural network component of the model architecture. As ResNet50 is solely responsible for extracting spatial features within this model architecture, rather than image classification output, the fully connected layer and softmax layer of ResNet50 were removed. A new fully connected layer was devised to align with the input dimension of the LSTM. The LSTM is responsible for processing and capturing temporal relationships in the sequence data, and it outputs the predicted phenological stage (Fig. 2a). Furthermore, the sequential fusion model can be optimized in the LSTM network by comparing the output of the intermediate time nodes with the output of the final time node, given the strict chronological order of wheat phenological stages. With the exception of the initial phenological stage label, the final time node's phenological stage label output is consistently temporally subsequent to the intermediate time node's phenological stage label output.

Synchronous fusion

Synchronous fusion employs a three-dimensional convolutional neural network (3D CNN) model architecture to integrate spatiotemporal features. In contrast to two-dimensional convolutional neural networks, which are limited to the consideration of spatial information, three-dimensional convolutional neural networks are capable of simultaneously analyzing both spatial and temporal features [5]. The samples of the wheat phenological stage consist of image series containing both spatial and temporal information. The synchronous fusion of spatiotemporal features in wheat phenological stage image series is achieved through the employment of 3D convolutional operations. In this study, a three-dimensional version of ResNet50, designated as 3D-ResNet50, was selected as the architectural foundation for the three-dimensional convolutional neural network model. The structure of 3D-ResNet50 is analogous to that of 2D-ResNet50, employing residual connections to facilitate the training of deep networks (Fig. 2d). The application of dilation to various modules, including convolutional layers, enables the 3D-ResNet50 model to process input samples from image series and subsequently generate the predicted label for the phenological stage (Fig. 2b).

Parallel fusion

Parallel fusion employs a dual-stream network architecture to fuse spatiotemporal features. A dual-stream network is comprised of two parallel convolutional neural networks, each processing optical flow and RGB images separately [34]. The predicted label is obtained through feature layer fusion. By capturing both dynamic and static features in wheat phenological stages, the dual-stream network model achieves a parallel fusion of spatiotemporal features. In this study, two parallel ResNet50 networks were constructed as a dual-stream network. One ResNet50 network processes optical flow and describes the direction and speed of pixel motion in images. This represents pixel-level motion patterns between adjacent frames in the image series. The other ResNet50 network is tasked with processing the last RGB image in the image series, which represents the spatial features. The softmax layer of each ResNet50 network outputs a probability distribution for the classification categories. The distributions are then combined to obtain the predicted label for the phenological stage. (Fig. 2c).

Experiment and results

Experimental parameter settings

The experiments were conducted on a server equipped with 2 Intel® Xeon® CPUs, 7 NVIDIA® TESLA® A100 GPUs (each with 40 GB memory), 1 TB of memory, and running Ubuntu 20.04. Furthermore, all three deep learning model architectures employed the backpropagation algorithm [31] to optimize the network parameters. The Adam optimizer was selected as the optimization algorithm, which incorporates the attributes of adaptive learning rate and momentum, facilitating the equilibrium between the convergence velocity and the performance of the model [18]. The cross-entropy loss function was employed, as it is a commonly utilized approach for multi-class classification tasks. This function effectively measures the discrepancy between the model's output probability distribution and the observed labels, thereby enhancing the accuracy and performance of the model in fitting the data [6]. In addition to the choice of optimizer and loss function, dropout was introduced to prevent overfitting [35]. This method involves randomly dropping out some neurons' outputs from the hidden layers of the network, thereby reducing the complexity of the model and improving its generalization ability. This regularization method helps to improve the model's generalization ability to unknown data and enhances the model's robustness. The hyperparameters for network training

were set as follows: a batch size of 16, a learning rate of 0.0001, and a dropout rate of 0.3.

Performance evaluation

To assess the model's performance more objectively and efficiently, this study selected the confusion matrix, overall accuracy (*OA*), mean absolute error (*MAE*), F1-score (*F1*), kappa coefficient (*Kappa*), and the number of parameters as evaluation metrics. The confusion matrix, an N×N grid, was used to display the correlation between predicted and actual labels, enabling the assessment of model performance for each category. *OA*, *F1*, and *Kappa* were derived from true positive (*TP*), false negative (*FN*), false positive (*FP*), and true negative (*TN*) cases. When an image labeled as stage *i* was correctly classified as stage *i*, it was considered *TP*. If an image labeled as stage *i* was misclassified as another stage, it was considered *FN*. Conversely, if an image was predicted as stage *i* but actually belonged to a different stage, it was regarded as *FP*. *OA*, *F1* and *Kappa* were determined using formulas 1–5 and ranged from 0 to 1, with higher values indicating better performance.

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision_i = \frac{TP_i}{TP_i + FP_i} \tag{2}$$

$$Recall_i = \frac{TP_i}{TP_i + FN_i} \tag{3}$$

$$F1\,score_i = \frac{2 * Precision_i * Recall_i}{Precision_i + Recall_i} \tag{4}$$

$$\begin{aligned} p_0 &= OA \\ p_e &= \frac{(\mathrm{TP} + \mathrm{FP})^*(\mathrm{TP} + \mathrm{FN}) + (\mathrm{FP} + \mathrm{TN})^*(\mathrm{FN} + \mathrm{TN})}{(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})^2} \\ Kappa &= \frac{p_0 - p_e}{1 - p_e} \end{aligned} \tag{5}$$

MAE is the mean of the absolute differences between predicted labels and observed labels, which is employed to assess the extent of the discrepancy between predicted and observed labels.

$$MAE = \sum_{i=1}^{n} \hat{y} - y \tag{6}$$

where *i* represents the *i-th* class, *n* represents the total number of samples, $\hat{y}$ represents the predicted label, and *y* represents the observed label.

Experimental results

The experimental results indicate that the three different spatiotemporal feature fusion methods proposed in this study effectively improve the overall accuracy of wheat phenological detection (Table 2, Table 3, Fig. 3). In terms of model complexity, the synchronous and parallel fusion models exhibit a notable increase in complexity relative to the baseline, with a parameter count of 48.93 M and 57.87 M, respectively. In terms of overall accuracy, the sequential fusion and synchronous fusion methods achieved the highest accuracy. In particular, these two methods got accuracies of 0.935 and 0.928, respectively, on the high-resolution dataset. In contrast, the parallel fusion method exhibited only a modest improvement in accuracy, reaching 0.888. With regard to the classification accuracy of different phenological stages, all three fusion methods demonstrate the highest performance in detecting the maturity stage. The sequential fusion and parallel fusion methods exhibited the lowest performance in detecting the anthesis stage, while the synchronous fusion method demonstrated the lowest performance in detecting the booting stage.

In the meantime, datasets with different resolutions exhibit different classification accuracy. The high-resolution dataset consistently demonstrates good classification accuracy, while the low-resolution dataset performs poorly. The performance of the medium-resolution dataset is slightly inferior to that of the high-resolution dataset. The discrepancies in accuracy between the three methods and the high-resolution dataset are 0.021, 0.009, and 0.015, respectively. In comparison, the discrepancies in accuracy between the three methods and the low-resolution dataset are considerably higher. The discrepancies in accuracy between the three methods are 0.088, 0.065, and 0.041, respectively (Fig. 4).

The angle of data collection has a profound effect on the model's training. The sequential fusion method yielded model accuracies of 0.874, 0.890, and 0.848 on datasets collected from viewing angles of 20° to 40°, 40° to 60°, and 60° to 80°, respectively. In the synchronous fusion method, the model accuracies on datasets from these angles are 0.885, 0.893, and 0.853, respectively. In the parallel fusion method, the model accuracies on datasets from these angles are 0.855, 0.861,

Table 2 The quantitative comparison of different methods

Method	OA	MAE	F1	Kappa	Params(M)
Baseline (ResNet50)	0.882	0.173	0.876	0.862	26.40
Sequential fusion	0.935	0.069	0.936	0.924	33.88
Synchronous fusion	0.928	0.073	0.914	0.917	48.93
Parallel fusion	0.888	0.160	0.884	0.870	57.87

Table 3 The F1-score of different methods in deriving different phenological stages

Method	Emergence	Tillering	Jointing	Booting	Heading	Anthesis	Filling	Maturity
Baseline	0.868	0.882	0.883	0.847	0.902	0.788	0.885	0.952
Sequential fusion	0.904	0.909	0.945	0.947	0.924	0.902	0.962	0.995
Synchronous fusion	0.950	0.958	0.935	0.806	0.835	0.826	1.000	1.000
Parallel fusion	0.873	0.886	0.901	0.882	0.891	0.819	0.884	0.939

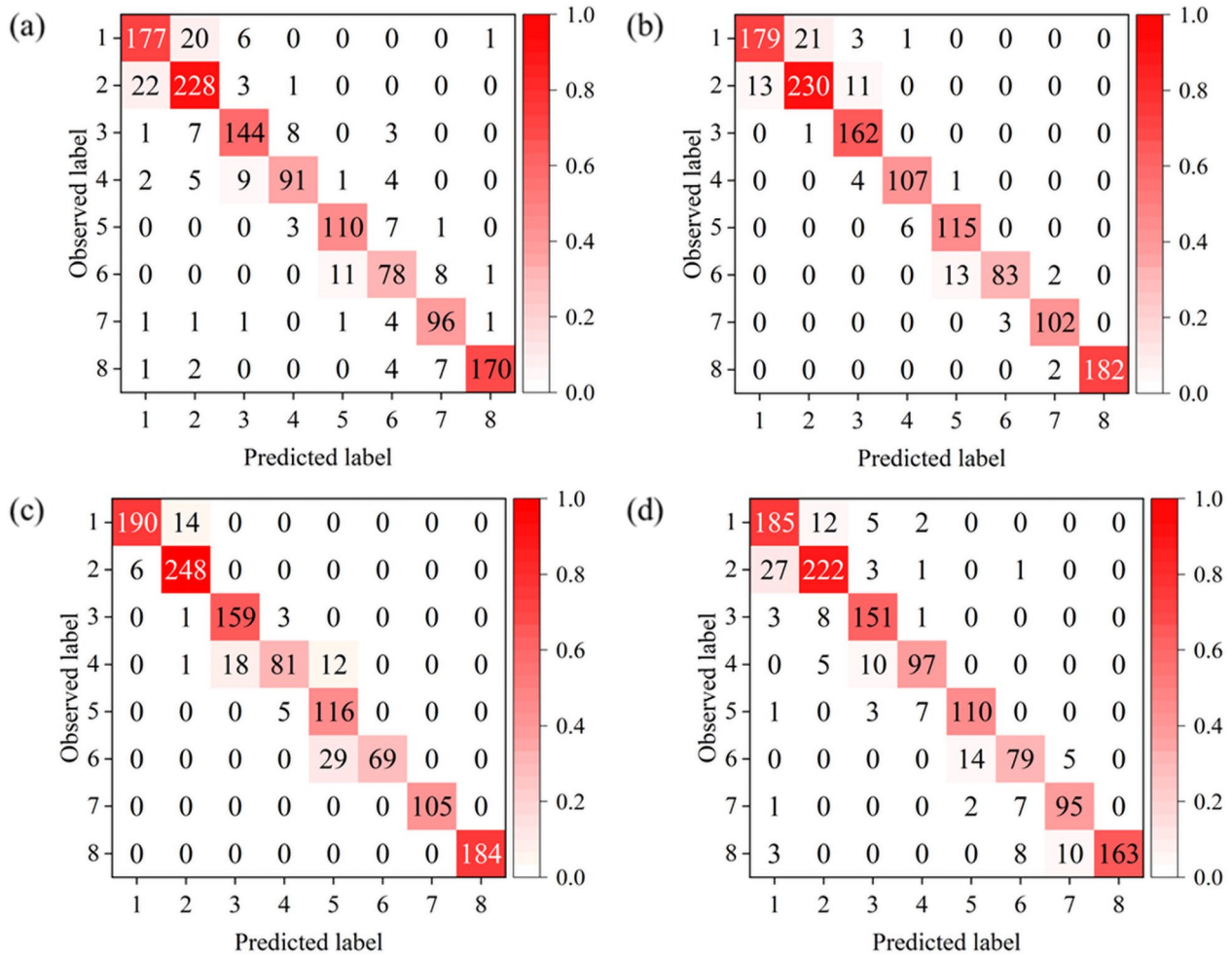

Fig.3 The confusion matrix yielded by four methods: baseline (**a**), sequential fusion (**b**), synchronous fusion (**c**), and parallel fusion (**d**). Serial numbers 1–8 represent the phenological stages: emergence, tillering, jointing, booting, heading, anthesis, filling, and maturity, respectively

and 0.834, respectively. The results demonstrate that when the acquisition angles are identical, synchronous fusion exhibits the most optimal performance, followed by sequential fusion. In contrast, parallel fusion demonstrates relatively inferior performance (Fig. 5). Across different fusion methods, models trained with data from different angles exhibit inferior performance compared to those trained with data from all angles. Nevertheless, within specific angle ranges, datasets ranging from 40° to 60° are best for deriving wheat phenology.

Moreover, different training strategies have a significant impact on the final accuracy of the models. In sequential fusion and synchronous fusion, fine-tuning training was found to improve the model accuracy by 0.086 and 0.070, respectively, in comparison to training

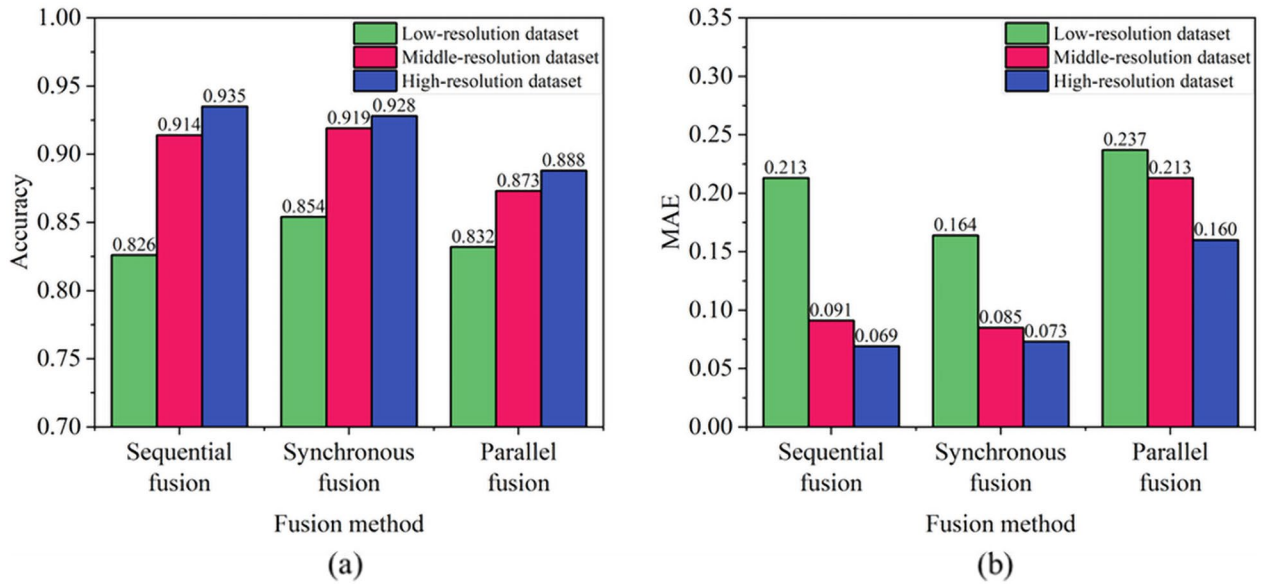

Fig.4 The performance of different fusion methods on different resolution datasets

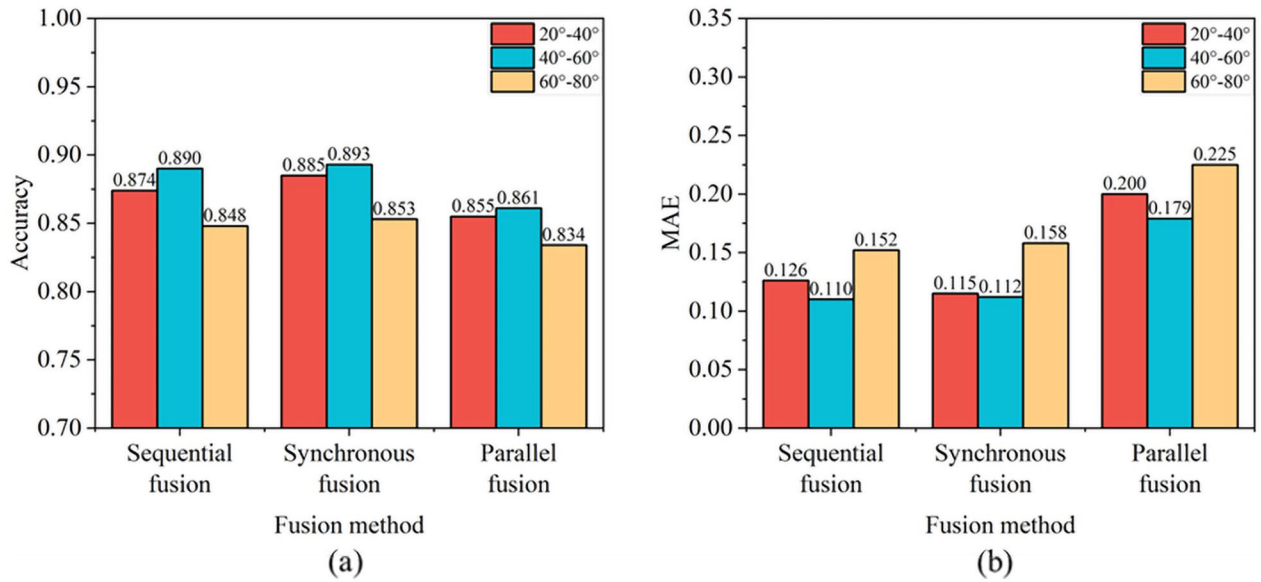

Fig.5 The performance of different fusion methods on datasets with different view angles. The specific angles of the acquired images are classified as 20° to 40°, 40° to 60°, and 60° to 80°. Datasets were divided based on these angles for the purpose of training

from scratch. In parallel fusion, fine-tuning training results in an improvement in model accuracy of 0.060 (Fig. 6).

Discussion

The three spatiotemporal feature fusion methods proposed in this study have effectively enhanced the performance of detecting wheat phenological stages. In contrast to traditional methods that solely employ single image inputs, which fail to consider the temporal aspects of wheat phenology [48, 52], this study employs image series as model inputs, integrating both spatial features and temporal features of the data, thereby enhancing the accuracy of phenological stages classification.

The three fusion methods proposed in this study are fundamentally different. Sequential fusion employs

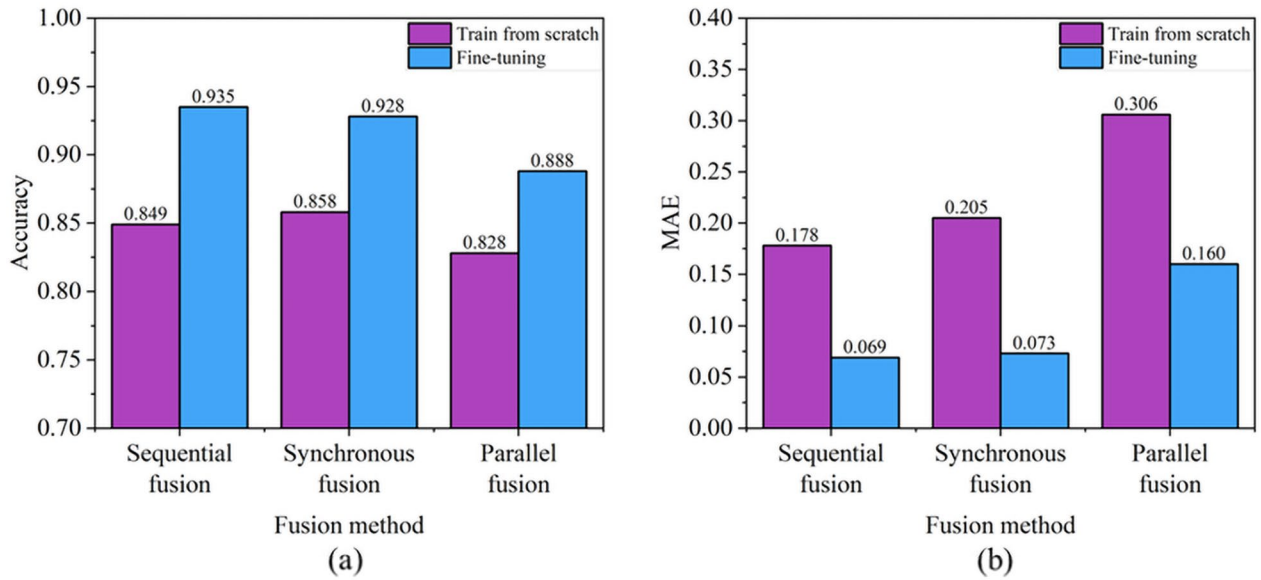

Fig.6 The performance of different fusion methods with different training strategies

ResNet50 to extract spatial features from input image series samples and generates feature maps in chronological order by using LSTM networks, which are employed to obtain classification results [2]. Synchronous fusion expands all two-dimensional operations in ResNet50 along the temporal dimension into three-dimensional operations, thereby enabling direct processing of image series samples. This approach achieves simultaneous feature extraction and classification of spatiotemporal features [12]. In parallel fusion, two ResNet50 networks are employed. One network is used for extracting spatial features from single images, while the other is used for extracting temporal features from optical flow series [36]. In the parallel fusion method, the optical flow features between image series are taken as temporal features of the samples. However, it should be noted that the image series samples in this study are not strictly continuous images. Consequently, while the outcomes achieved through the utilization of optical flow characteristics exhibit a marginal improvement in comparison to training with single-stage images, the enhancement in accuracy is relatively modest, and the impact is not pronounced.

In contrast, both sequential fusion and synchronous fusion demonstrate excellent performance in extracting temporal features, resulting in a notable enhancement in accuracy. Indeed, sequential fusion entails the gradual incorporation of an LSTM network into the framework of a convolutional neural network, thereby facilitating a progressive integration of information [30]. In contrast, synchronous fusion involves extending two-dimensional operations to three-dimensional operations during the feature extraction process, thereby enabling comprehensive and synchronized feature fusion at each step [13]. Consequently, the synchronous fusion method exhibits a greater parameter count compared to the sequential fusion method, resulting in increased consumption of memory resources and runtime during both model training and prediction. Furthermore, the sequential fusion method enables the optimization and adjustment of output time nodes based on prior knowledge. The optimized sequential fusion method achieved an accuracy of 0.935, which was higher than that of the synchronous fusion method (Fig. 7).

The diverse deployment methods of near-surface cameras result in a corresponding diversity in the resolution and angles at which images are captured. The slow change in the appearance of wheat phenology makes it challenging to distinguish between images of adjacent stages in the task of wheat phenological stages classification [59]. The inclusion of more spatial information in high-resolution images leads to enhanced performance in the detection of wheat phenological stages [53]. The results of the study (Fig. 4) indicate that an increase in image resolution significantly enhances the performance of classification, particularly in stages involving subtle features. The results demonstrated that high-resolution datasets exhibited superior performance, while low-resolution datasets exhibited significantly poorer performance. The performance of the medium-resolution dataset is intermediate between the two extremes, exhibiting a significant improvement compared to the

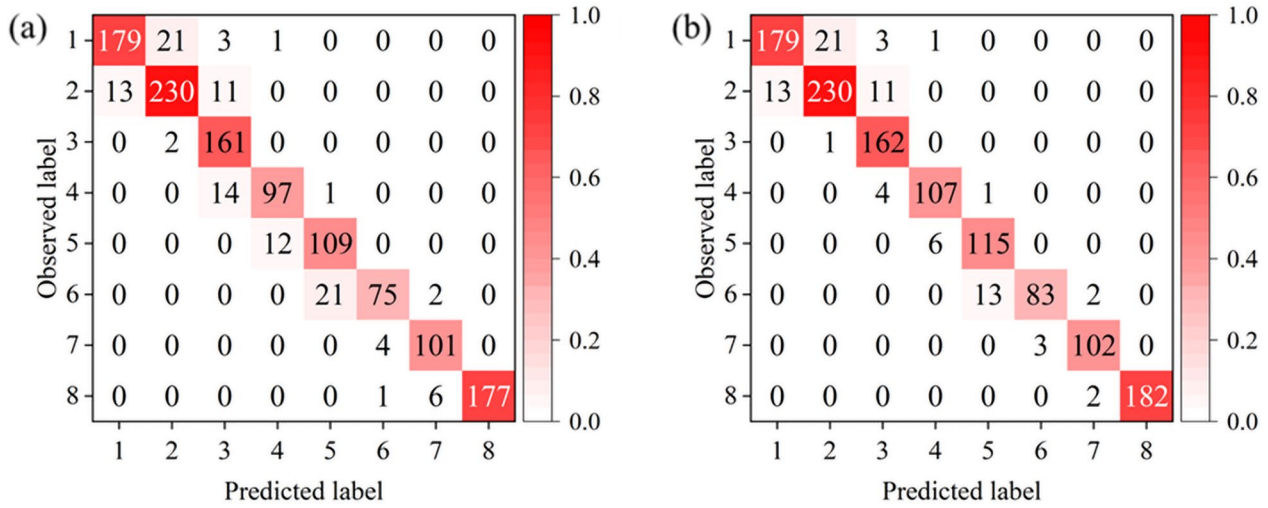

Fig.7 The confusion matrix yielded by sequential fusion without prior knowledge (**a**) and sequential fusion with prior knowledge (**b**)

low-resolution dataset and a relatively minor difference compared to the high-resolution dataset. Consequently, this study indicates that both medium-resolution and high-resolution datasets can effectively accomplish the task of wheat phenological stages classification. However, the optimal choice between the two should be based on the memory resources available during training.

In the meantime, the outcomes of the wheat phenological stages classification at various image acquisition angles demonstrate that datasets within the vertical range of 40° to 60° are best in this study (Fig. 5). This phenomenon may be attributed to the fact that the shooting angle within this range allows for the capture of a greater number of wheat plants, thereby reducing the uncertainty associated with heterogeneity. In the meantime, images captured at greater angles encompass features of varying scales, furnishing the model with a more comprehensive array of information for the extraction of features representing different phenological stages [11]. The data within the range of 20° to 40° provide information at the organ scale, such as wheat type and color [3]. In contrast, the dataset within the range of 40° to 60° includes a more significant number of phenological features, such as the proportion of spikes, stems, and leaves, as well as the curvature of spikes and the collective information of the canopy [52]. However, data within the range of 60° to 80° may be less detailed due to occlusion by wheat leaves in the foreground. Consequently, when utilizing near-surface cameras at a height of 3 m, the angle range of 40° to 60° is deemed to offer the most advantageous outcome.

Moreover, the performance of deep learning models is significantly influenced by the training strategies employed. Training a deep learning network from scratch necessitates the availability of a substantial quantity of annotated training data, such as those found in the ImageNet [8] or COCO [23] datasets. However, in the field of crop phenology, the acquisition of large-scale publicly available datasets is challenging [55]. Consequently, in the absence of millions of labeled data to support it, training a network from scratch is not the optimal approach for optimizing model parameters [37]. Given that the three fusion methods employed in this study are all based on the same baseline, we employed a pre-trained ResNet50 as an initial model. In both sequential fusion and parallel fusion, the pre-trained ResNet50 was employed directly. However, in synchronous fusion, each layer of the ResNet50 is expanded from a two-dimensional structure to a three-dimensional structure. Implementing corresponding expansion operations on the pre-trained model is necessary. Specifically, convolutional kernels are copied along the temporal dimension and evenly distributed across all temporal dimensions. Each three-dimensional convolution contains the same pretrained parameters across all temporal dimensions, thus ensuring that the input dimensions of the three-dimensional convolutional layer are matched [49]. The results indicate that the utilization of pre-trained models and the subsequent fine-tuning of the model parameters through backpropagation can further enhance model performance (Fig. 6).

Previous studies have demonstrated that soil background [20] and weather conditions [28] can significantly influence the accuracy of wheat phenological stage detection. For example, soil under sunlight may produce considerable shadows [27], while rainy conditions might introduce water reflections that affect image quality [45]. To address these challenges, our study employed a broader time range for data collection, extending from 8:00 to 17:00 daily and encompassing a

variety of weather conditions, including cloudy, sunny, and rainy days. This approach not only enhances the generalization capability of the dataset but also enables the model to perform effectively across diverse daytime field environments. This method leverages the advantages of near-surface cameras in the field, which can capture images at any time of day, unrestricted by weather conditions. However, it should be noted that our dataset is limited to RGB images captured during daylight hours, excluding nighttime data, which presents a limitation of the current study. Future research could address this by incorporating near-infrared data collected during nighttime [9, 26, 47], thereby further improving the model's applicability and generalization capacity. Additionally, while the sequential fusion architecture, as described in this study, has been demonstrated to be effective in the task of wheat phenological stages classification. However, it should be noted that this approach also increases the model's parameter count. Future research will concentrate on the reduction of model complexity and the optimization of resource requirements through the application of model compression and simplification. Among the techniques mentioned above, knowledge distillation [10], weight quantization [1], and lightweight model design [46] will be of particular importance. Implementing these technologies helps simplify model structures and reduce computational and storage requirements without sacrificing model performance. The above methods allow the models to be made lighter, making them easier to apply to practical near-surface camera systems. This lightweight model structure will provide more convenient and efficient solutions for real-time monitoring and decision support in the agricultural sector, providing strong support for the intelligent development of agricultural production.

Conclusion

This study proposes a new approach for deriving wheat phenological stages based on near-surface RGB image series and three different spatiotemporal feature fusion methods. The results indicate that the sequential fusion architecture is an effective method for detecting the phenological stages of wheat and achieves a balance between performance and resource consumption. Furthermore, the employment of high-resolution datasets, two-stage fine-tuning training, and observations within the 40° to 60° range can also enhance the performance of the model. The findings of this study have broader implications, as the methodology developed here holds the potential to be extended to other crops in future research. By enabling more accurate monitoring of the phenological stages across various crops, this approach provides a robust foundation for optimizing agricultural practices, improving crop management, and ultimately contributing to a more sustainable agricultural system. The ability to precisely monitor crop phenology is critical for resource-efficient agriculture, and this research represents a significant advancement in that direction.

Author contribution

Y.C. performed experiments, analyzed the data, and wrote the manuscript. Y.L. performed experiments, analyzed the data, and wrote the manuscript. X.Q. performed experiments and prepared data visualization. J.Z. performed experiments. L.J. developed software used in this work. Y.T. supervised the research activity planning and execution. Y.Z. supervised the research activity planning and execution. W.C. managed and coordinated the research activity planning and execution. X.Z. conceived the research, guided the entire study, revised the manuscript, and provided valuable comments and suggestions. All the authors approved the manuscript and have made all required statements and declarations.

Funding

This research was supported by the National Natural Science Foundation of China (Grant No. 32171892), the Qing Lan Project of Jiangsu Universities and Jiangsu Agricultural Science and Technology Innovation Fund (CX (21) 1006).

Availability of data and materials

No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]National Engineering and Technology Center for Information Agriculture, Nanjing Agricultural University, Nanjing 210095, China. [2]Key Laboratory for Crop System Analysis and Decision Making, Ministry of Agriculture and Rural Affairs, Nanjing 210095, China. [3]College of Geography, Jiangsu Second Normal University, Nanjing 211200, China. [4]College of Agricultural Engineering, Jiangsu University, Zhenjiang 212013, China. [5]Jiangsu Collaborative Innovation Center for Modern Crop Production, Nanjing 210095, China.

Received: 11 June 2024 Accepted: 23 September 2024
Published online: 30 September 2024

References

1. Aaron A, Hassan M, Hamada M, et al. A lightweight deep learning model for identifying weeds in corn and soybean using quantization. Eng Proc. 2023;56(1):318.
2. Bai X, Xue R, Wang L, et al. Sequence SAR image classification based on bidirectional convolution-recurrent network. IEEE Trans Geosci Remote Sens. 2019;57(11):9223–35.
3. Bekkering CS, Huang J, Tian L. Image-based, organ-level plant phenotyping for wheat improvement. Agronomy. 2020;10(9):1287.
4. Cardona J, Howland M, Dabiri J. Seeing the wind: Visual wind speed prediction with a coupled convolutional and recurrent neural network. Adv Neural Inform Proc Syst, 32. 2019.
5. Carreira J, Zisserman A. Quo vadis, action recognition? a new model and the kinetics dataset. Proc IEEE Conf Comput Vision Pattern Recogn 2017.

6. De Boer P-T, Kroese DP, Mannor S, et al. A tutorial on the cross-entropy method. Ann Oper Res. 2005;134:19–67.
7. Delavarpour N, Koparan C, Nowatzki J, et al. A technical study on UAV characteristics for precision agriculture applications and associated practical challenges. Remote Sens. 2021;13(6):1204.
8. Deng J, Dong W, Socher R, et al. Imagenet: a large-scale hierarchical image database. 2009 IEEE Conf Comput Vision Pattern Recogn 2009. IEEE.
9. Fan X, Kawamura K, Guo W, et al. A simple visible and near-infrared (V-NIR) camera system for monitoring the leaf area index and growth stage of Italian ryegrass. Comput Electron Agric. 2018;144:314–23.
10. Ghofrani A, Mahdian TR. Knowledge distillation in plant disease recognition. Neural Comput Appl. 2022;34(17):14287–96.
11. Han J, Shi L, Yang Q, et al. Real-time detection of rice phenology through convolutional neural network using handheld camera images. Precision Agric. 2020;22(1):154–78.
12. Hara K, Kataoka H, Satoh Y. Learning spatio-temporal features with 3D residual networks for action recognition. Proc IEEE Int Conf Comput Vision Workshops 2017.
13. Hara K, Kataoka H, Satoh Y. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet?. Proc IEEE Conf Comput Vision Pattern Recogn 2018.
14. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016.
15. Islam MM, Talukder MA, Sarker MRA, et al. A deep learning model for cotton disease prediction using fine-tuning with smart web application in agriculture. Intell Syst Appl. 2023;20:200278.
16. Jia D, Cheng C, Song C, et al. A hybrid deep learning-based spatiotemporal fusion method for combining satellite images with different resolutions. Remote Sens. 2021;13(4):645.
17. Kamilaris A, Prenafeta-Boldú FX. Deep learning in agriculture: a survey. Comput Electron Agricu. 2018;147:70–90.
18. Kingma D P, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv: 14126980, 2014.
19. Li D, Chen JM, Yu W, et al. Assessing a soil-removed semi-empirical model for estimating leaf chlorophyll content. Remote Sens Environ. 2022;282: 113284.
20. Li L, Minzan L, Gang L, et al. Goals, key technologies, and regional models of smart farming for field crops in China. Smart Agricu. 2022;4(4):26–34.
21. Li X, Hou B, Zhang R, et al. A review of RGB image-based internet of things in smart agriculture. IEEE Sens J. 2023. https://doi.org/10.1109/JSEN.2023.3309774.
22. Liao C, Wang J, Shan B, et al. Near real-time detection and forecasting of within-field phenology of winter wheat and corn using sentinel-2 time-series data. ISPRS J Photogram Remote Sens. 2023;196:105–19.
23. Lin T-Y, Maire M, Belongie S, et al. Microsoft coco: common objects in context. computer vision–ECCV 13th European conference, Zurich, Switzerland, September 6–12, Proceedings, Part V 13. Berlin: Springer; 2014.
24. Liu S, Jin S, Guo Q, et al. An algorithm for estimating field wheat canopy light interception based on digital plant phenotyping platform. Smart Agricu. 2020;2(1):87.
25. Liu S, Peng D, Zhang B, et al. The accuracy of winter wheat identification at different growth stages using remote sensing. Remote Sens. 2022;14(4):893.
26. Lu M, Wang H, Xu J, et al. A Vis/NIRS device for evaluating leaf nitrogen content using K-means algorithm and feature extraction methods. Comput Electron Agric. 2024;225: 109301.
27. Marais-Sicre C, Queguiner S, Bustillo V, et al. Sun/shade separation in optical and thermal UAV images for assessing the impact of agricultural practices. Remote Sens. 2024;16(8):1436.
28. Osipov A, Pleshakova E, Gataullin S, et al. Deep learning method for recognition and classification of images from video recorders in difficult weather conditions. Sustainability. 2022;14(4):2420.
29. Patrício DI, Rieder R. Computer vision and artificial intelligence in precision agriculture for grain crops: a systematic review. Comput Electr Agricu. 2018;153:69–81.
30. Rani CJ, Devarakonda N. An effectual classical dance pose estimation and classification system employing convolution neural network–long short term memory (CNN-LSTM) network for video sequences. Microproc Microsyst. 2022;95: 104651.
31. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986;323(6088):533–6.
32. Ruml M, Vulić T. Importance of phenological observations and predictions in agriculture. J Agricu Sci. 2005;50(2):217–25.
33. Sharma A, Jain A, Gupta P, et al. Machine learning applications for precision agriculture: a comprehensive review. IEEE Access. 2020;9:4843–73.
34. Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. Adv Neural Inform Proc Syst. 27. 2014
35. Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting. J Machine Learn Res. 2014;15(1):1929–58.
36. Sun S, Kuang Z, Sheng L, et al. Optical flow guided feature: a fast and robust motion representation for video action recognition. Proc IEEE Conf Comput Vision Pattern Recogn 2018.
37. Tajbakhsh N, Shin JY, Gurudu SR, et al. Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Trans Med Imaging. 2016;35(5):1299–312.
38. Taylor SD, Browning DM. Classification of daily crop phenology in phenocams using deep learning and hidden Markov models. Remote Sens. 2022. https://doi.org/10.3390/rs14020286.
39. Too EC, Yujian L, Njuki S, et al. A comparative study of fine-tuning deep learning models for plant disease identification. Comput Electron Agric. 2019;161:272–9.
40. Wang S, Zhao J, Cai Y, et al. A method for small-sized wheat seedlings detection: from annotation mode to model construction. Plant Methods. 2024;20(1):15.
41. Wang X, Girshick R, Gupta A, et al. Non-local neural networks. Proc IEEE Conf Comput Vision Pattern Recogn, 2018.
42. Wang XA, Tang J, Whitty M. DeepPhenology: estimation of apple flower phenology distributions based on deep learning. Comput Elect Agricu. 2021;185: 106123.
43. Wang Y, Zhang X, Ma G, et al. Recognition of weeds at asparagus fields using multi-feature fusion and backpropagation neural network. Int J Agricu Biol Eng. 2021;14(4):190–8.
44. Wei L, Yang H, Niu Y, et al. Wheat biomass, yield, and straw-grain ratio estimation from multi-temporal UAV-based RGB and multispectral images. Biosys Eng. 2023;234:187–205.
45. Wu Z, Wang Z, Spohrer K, et al. Non-contact leaf wetness measurement with laser-induced light reflection and RGB imaging. Biosys Eng. 2024;244:42–52.
46. Xie Y, Zhong X, Zhan J, et al. ECLPOD: an extremely compressed lightweight model for pear object detection in smart agriculture. Agronomy. 2023;13(7):1891.
47. Xiong Y, McCarthy C, Humpal J, et al. Near-infrared spectroscopy and deep neural networks for early common root rot detection in wheat from multi-season trials. Agron J. 2024. https://doi.org/10.1002/agj2.21648.
48. Yalcin H. Plant phenology recognition using deep learning: deep-pheno. 2017 6th Int Conf Agro-Geoinform. 2017. https://doi.org/10.1109/Agro-Geoinformatics.2017.8046996.
49. Yang J, Huang X, He Y, et al. Reinventing 2D convolutions for 3D images. IEEE J Biomed Health Inform. 2021;25(8):3009–18.
50. Yang N, Yuan M, Wang P, et al. Tea diseases detection based on fast infrared thermal image processing technology. J Sci Food Agricu. 2019;99(7):3459–66.
51. Yang Q, Shi L, Han J, et al. A near real-time deep learning approach for detecting rice phenology based on UAV images. Agricu Forest Meteorol. 2020. https://doi.org/10.1016/j.agrformet.2020.107938.
52. Yang Z, Gao S, Xiao F, et al. Leaf to panicle ratio (LPR): A new physiological trait indicative of source and sink relation in japonica rice based on deep learning. Plant Methods. 2020;16:1–15.
53. Zhang C, Marzougui A, Sankaran S. High-resolution satellite imagery applications in crop phenotyping: an overview. Comput Electr Agricu. 2020;175: 105584.
54. Zhang J, Yang C, Song H, et al. Evaluation of an airborne remote sensing platform consisting of two consumer-grade cameras for crop identification. Remote Sens. 2016;8(3):257.
55. Zhang R, Jin S, Zhang Y, et al. PhenoNet: a two-stage lightweight deep learning framework for real-time wheat phenophase classification. ISPRS J Photogramm Remote Sens. 2024;208:136–57.

56. Zhao J, Cai Y, Wang S, et al. Small and oriented wheat spike detection at the filling and maturity stages based on WheatNet. Plant Phenomics. 2023;5:0109.
57. Zhao J, Yan J, Xue T, et al. A deep learning method for oriented and small wheat spike detection (OSWSDet) in UAV images. Comput Electr Agricu. 2022;198: 107087.
58. Zhou M, Ma X, Wang K, et al. Detection of phenology using an improved shape model on time-series vegetation index in wheat. Comput Elect Agricu. 2020. https://doi.org/10.1016/j.compag.2020.105398.
59. Zhou Q, Guo W, Chen N, et al. Analyzing nitrogen effects on rice panicle development by panicle detection and time-series tracking. Plant Phenomics. 2023;5:0048.

Publisher's Note