Precision Agriculture (2021) 22:154–178
https://doi.org/10.1007/s11119-020-09734-2

Real-time detection of rice phenology through convolutional neural network using handheld camera images

Jingye Han[1] · Liangsheng Shi[1] · Qi Yang[1] · Kai Huang[2] · Yuanyuan Zha[1] · Jin Yu[1]

Published online: 28 June 2020

Abstract
Smallholder farmers play an important role in the global food supply. As smartphones become increasingly pervasive, they enable smallholder farmers to collect images at very low cost. In this study, an efficient deep convolutional neural network (DCNN) architecture was proposed to detect development stages (DVS) of paddy rice using photographs taken by a handheld camera. The DCNN model was trained with different strategies and compared against the traditional time series Green chromatic coordinate (time-series Gcc) method and the manually extracted feature-combining support vector machine (MF-SVM) method. Furthermore, images taken at different view angles, model training strategies, and interpretations of predictions of the DCNN models were investigated. Optimal results were obtained by the DCNN model trained with the proposed two-step fine-tuning strategy, with a high overall accuracy of 0.913 and low mean absolute error of 0.090. The results indicated that images taken at large view angles contained more valuable information and the performance of the model can be further improved by using images taken at multiple angles. The two-step fine-tuning strategy greatly improved the model robustness against the randomness of view angle. The interpretation results demonstrated that it is possible to extract phenology-related features from images. This study provides a phenology detection approach to utilize handheld camera images in real time and some important insights into the use of deep learning in real world scenarios.

Keywords Phenology · Rice · CNN · Deep learning · Handheld camera image

Introduction

Smallholder farmers, defined generally as those who own plots smaller than 2 ha, are estimated to produce 30–50% of the global food supply (Ricciardi et al. 2018), and account for over 80% of farms worldwide (Lowder et al. 2016). In regions where smallholder farming dominates the agricultural landscape—for example, in sub-Saharan Africa, India, and

✉ Liangsheng Shi
liangshs@whu.edu.cn

[1] State Key Laboratory of Water Resources and Hydropower Engineering Sciences, Wuhan University, Wuhan 430072, Hubei, China

[2] Guangxi Hydraulic Research Institute, Nanning 530023, Guangxi, China

China—food security and sustainability depend on how smallholders farm their land (Cui et al. 2018). Precision agriculture comprises a set of technologies that combine sensors, information systems, and informed management to optimize production (Gebbers and Adamchuk 2010). One method of implementing precision agriculture is based on real-time and accurate information on crop growth and the relevant environmental conditions (Zhang et al. 2002). As smartphones become cheaper and ubiquitous, they can be used as sensors to help smallholders, who typically have limited resources and knowledge, access precision agriculture technologies.

For efficient crop management, phenology information is essential to meet the right dates for irrigation, fertilizing, and crop protection (Schwartz 2013; Jamieson et al. 2007; Zheng et al. 2016). According to the sensor platform used, methods of phenology monitoring can be divided into three groups: (1) satellite platforms with low temporal frequency (from 12 h to 10 days) time-series data at a global scale using sensors with a coarse resolution (from 250 m to 1 km), such as Moderate Resolution Imaging Spectroradiometer (MODIS) (Huete et al. 2013); (2) unmanned aerial vehicle platforms equipped with digital cameras and multispectral sensors to collect images and vegetation index at a regional scale at high temporal (every day) and spatial (from 1 cm to 1 m) resolutions (Klosterman et al. 2018; Park et al. 2019); and (3) near-surface platforms with digital cameras and thermal infrared cameras that continuously (every hour) acquire images at an ultra-high spatial resolution (from 1 mm to 1 cm) (Melaas et al. 2018; Petach et al. 2014; Sonnentag et al. 2012).

Satellite remote sensing is the most common method. Although it provides image covering a wide range, its lack of spatial and temporal resolution, in general, makes it unsuitable for the continuous monitoring of plant phenology (White et al. 2009; Zhang et al. 2006). Moreover, UAVs are unsuitable for collecting data in rainy or windy weather. A given phenological stage of paddy rice, such as heading or anthesis, takes 5 to 14 days to complete (Yoshida 1981), which means that when rain lasts longer than a week, satellite data (e.g., MODIS eight-day product) and UAV data cannot be used in time to detect key stages of growth.

To address the above problems, a near-surface remote sensing method has been developed to monitor the growth of vegetation from the organ to the landscape scale (Richardson et al. 2009; Putra and Soni 2019; Sunoj et al. 2016). Sakamoto et al. (2012) estimated maize biophysical characteristics using digital photographs. Guo et al. (2015) used field-acquired time-series (every 5 min from 8:00 to 16:00) RGB images to automatically characterize the dynamics of flowering in the anthesis stage of rice. Bai et al. (2018) used a fixed camera to detect rice spikes and proposed a method for the automatic observation of the heading stage. In addition, the Phenocam network, a network consisting of dedicated surveillance digital cameras that capture photographs of the plant canopy at a desired time interval over the duration of plant growth, provided a series of images for studying the phenological impacts of climate change (Petach et al. 2014; Sonnentag et al. 2012; Sunoj et al. 2016; Zhang et al. 2018).

However, owing to the significant differences in features among the development stages, traditional methods of identifying crop phenology based on near-surface remote sensing focus on specific stages, such as the emergence and three-leaf stages of maize (Yu et al. 2013), heading stages of wheat and rice (Hufkens et al. 2018; Zhu et al. 2016), and anthesis stage of rice (Guo et al. 2015). Most of these methods require images taken by a fixed camera, which are not suitable for smallholders who have many scattered fields. To overcome this shortcoming, Hufkens et al. (2018) assessed the capability of phenocam-style time-series data collection to support phenology monitoring in agriculture and found it can be used to quantify the

development stages. But the demand for time-series data made this method difficult to apply to real-time estimation. For the convenience of practical application, a robust method is still required to detect all development stages of crops using ordinary handheld camera images.

In recent years, significant advances in data collection techniques and computing resources led to a boom of deep learning (DL). The application of DL to agriculture by using UAV imagery and near-surface photograph has been reviewed by Kamilaris and Prenafeta-Boldú (2018) and Patrício and Rieder (2018). Among all DL methods, the deep convolutional neural network (DCNN) exhibits impressive performance on image classification (Krizhevsky et al. 2012) and regression tasks (Liu et al. 2015). For the application of the convolutional neural network to precision agriculture, considerable attention has been paid to disease detection (Ferentinos 2018; Sladojevic et al. 2016), fruit or ear counting (Chen et al. 2017; Koirala et al. 2019; Liu et al. 2018; Madec et al. 2019; Stein et al. 2016), weed detection (Milioto et al. 2018; Sa et al. 2018), and crop segmentation (Dyson et al. 2019). Yang et al. (2019) proposed the use of DCNNs for yield prediction ($R^2=0.585$) and Ma et al. (2019) proposed a network to estimate above ground biomass ($R^2=0.808$).

The hypothesis of this study is that the features of the crop phenotype can be captured by machine learning through analyzing images, while they are traditionally recognized by agricultural expert through observation. However, deep learning studies for crop phenology detection are still very limited. Yalcin (2017) applied the DCNN to classify the development stages by using fixed-angle images. Bai et al. (2018) used the support vector machine and DCNN to distinguish the image patches of rice spike. The number of spike patches detected determined the rice heading stage. The above two studies focused on images at a fixed view angle and location, while the smallholder famers may take images at a random view angle and position. It is necessary to develop a versatile method for the convenience of dealing with these random images. It will be attractive to extract maximum phenology information from images taken at multiple angles. A training strategy is also required to enhance the performance of deep learning method by lessening the influence of the uncertainty of image view angle.

Three contributions are made in this study. First, a method which can identify rice phenology using handheld camera image is proposed in contrast to most studies that have employed satellite and UAV remote sensing data or RGB photos at a fixed view angle and location. Second, real-time images are used to identify rice phenology. This is superior to approaches based on time-series data, where the phenology can be identified only after data for the entire growth season have been collected. Third, all development stages were detected by a DCNN model, whereas most previous studies have focused on single-stage identification. The objectives of this study are to (1) develop a new DCNN architecture to identify phenology using handheld camera RGB photos and to simultaneously utilize multi-angle images for maximum utilization of different images; (2) develop a new training strategy to improve the model robustness against the randomness of view angle caused by handheld shooting; and (3) attempt to interpret predictions of deep learning method for phenology detection of paddy rice by visualizing the result of predictions.

Study area

The experimental site (23° 5′ 52″–23° 7′ 23″ N, 108° 57′ 7″–108° 58′ 34″ E) was located in Binyang county of Guangxi province of China (Fig. 1). The 160 ha of the area were divided into more than 800 plots managed by local farmers. The yearly average precipitation over this region was approximately 1600 mm with an average temperature of 21 °C.

A total of 70 plots owned by different holders were randomly selected and 12 plots managed by us were used for analysis. The rice seedlings were transplanted from 22 July to 15 August 2018 and were harvested from 2 November to 25 November 2018.

Data collection and processing

Image acquisition

The handheld camera RGB images were taken using a handheld digital camera. A QX-1 (SONY, Japan) was used for the first time and the remaining images were acquired by a DSC-RX1RM2 (SONY, Japan). The RGB sensors had 7952 × 5304 pixels and 5456 × 3632 pixels for the DSC-RX1RM2 and the QX-1, respectively. The camera was operated with automatic exposure control and exposure compensation was employed when the illumination was insufficient. Most of the images were taken between 9:00 and 17:00. Due to the large number of plots, it was difficult to control the consistent ambient light when taking images during this period. Therefore, the images taken at different times did have differences in ambient light, but the effect of ambient light could be reduced by data augmentation scheme in "Data augmentation". To utilize the images taken at different view angles for phenology identification, four vertical directions—0° (A), 20° (B), 40° (C), and 60° (D)—between the direction of photography and that of gravity were chosen (Fig. 2a). In the study area, most plots were transplanted by drill planter. When images were taken in the sowing direction, the soil between the two rows of rice would be well captured, while in other directions, the soil was less captured by images. Three horizontal directions—0° (a), 45° (b), and 90° (c)—between the directions of photography and the sowing direction were thus set to avoid the effects of drilling (Fig. 2b). Twelve photos were taken in each observation for a plot, and the view angle was roughly controlled by hand at 1.5 m

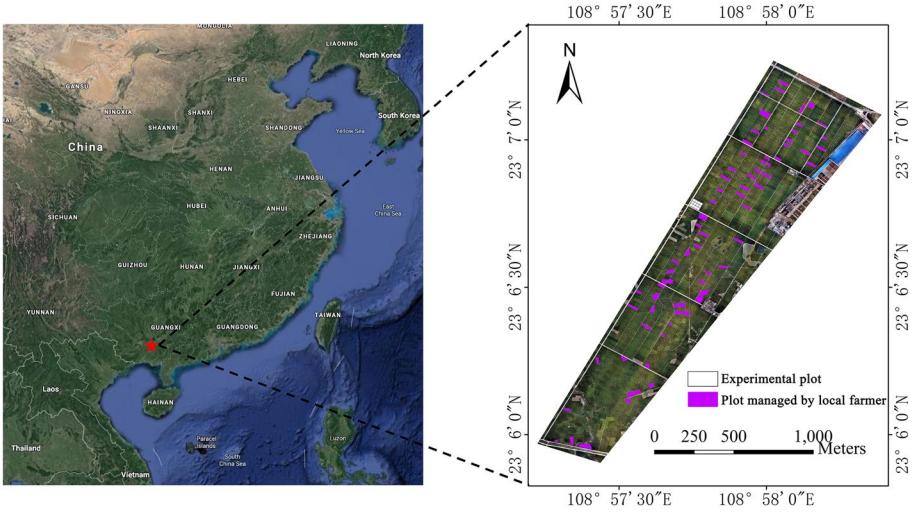

Fig. 1 Study area: RGB orthostatic map on 13 September 2018 (right) of summer-autumn rice experiment in Guangxi province, China. Management units of 12 experimental plots with development stage record are displayed as white polygons and the 70 plots managed by local farmers are displayed as purple polygons (Color figure online)

above the ground (Fig. 2a). Image collection was deployed seven times for 70 plots managed by local farmers (Fig. 1) and 11 times for 12 experimental plots (Fig. 1). A dataset with 622 observations (70 plots × 7 times + 12 plots × 11 times) containing 7 464 images (622 × 12 angles) was constructed, and 7 320 (610 × 12 angles) of them were used for analysis. The other images were poor in quality and could not be used. The 610 observations were divided into 10 groups according to the DVS, and each group was divided further into training (60%), validation (20%), and testing (20%) sets.

Field data acquisition

The field DVS observations and image acquisition were conducted at the same time. Five clusters were chosen randomly from a plot, and they were further classified according to the principal code for the development stages of the BBCH (Lancashire et al. 1991). Averages of the code for the development stages based on the five clusters were calculated as true stages.

Data augmentation

A DCNN trained by the original dataset tends to overfit because labeled samples are scarce (Perez and Wang 2017). Therefore, a data augmentation scheme was used to increase the size of the image dataset and reduce chances of overfitting (Fig. 3).

First, the image dataset was cropped by a square region with a side length 8/9th of the height of the image. Owing to the imbalanced data distribution (Fig. 5a), a cropping scheme (Table 1) was used to balance the training and validation sets. The photos in the test set were cropped six times. Figure 4 shows a cropped patch from the original image. After cropping, the resulting distribution of the dataset was more even than the original distribution (Fig. 5b).

Second, a horizontal flip was used to double the size of the dataset. The datasets were then further augmented by simulating changes in illumination change by transferring the images to HSV color space and adjusting the values of V (Smith 1978). The value was increased and decreased by 20%, respectively, to triple the training and the validation sets. No flip or HSV adjustment was applied to photos in the test set.

Fig. 2 Image acquisition by the handheld camera: **a** four photos were taken at the vertical angles of 0°, 20°, 40° and 60° and at the height of 1.5 m; **b** three horizontal directions were chosen to take the photos

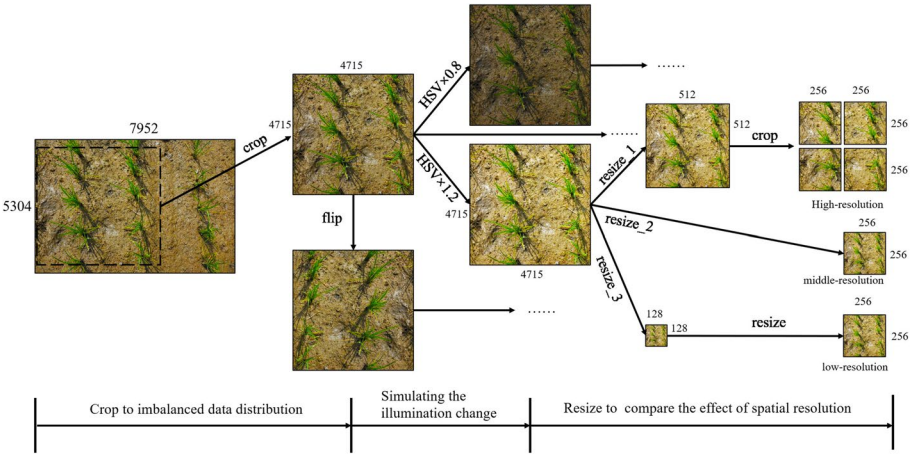

Fig. 3 The data augmentation scheme. The original photo with the size of 7952×5304 is taken by DSC-RX1RM2. As for QX-1, the size of the original photo is 5456×3632, and the size of the image cropped from it is 3228×3228. The ellipsis indicates that the image will be processed the same as other images

Table 1 The cropping scheme of three datasets

DVS class	Training and validation set			Test set		
	Number of X	Number of Y	Crop time	Number of X	Number of Y	Crop time
0	11	10	110	3	2	6
1	6	3	18			
2	3	1	3			
3	2	2	4			
4	3	2	6			
5	8	5	40			
6	5	3	15			
7	3	2	6			
8	3	1	3			
9	10	8	80			

X and Y are the abscissa and ordinate of point O in Fig. 4

Finally, images in the three sets were resized to three sizes: 128×128, 256×256, and 512×512. The 128×128 images were resized again to 256×256, and the 512×512 images were divided into four 256×256 images. These three datasets with low, middle, and high resolutions were used to compare the effects of spatial resolution on the DCNN.

Methods

DCNNs were used to identify the DVS of rice, and their performance was compared with that of the Gcc-time-series (Melaas et al. 2018) and manually extracted feature-combining SVM approaches (Yalcin 2017; Ma et al. 2019).

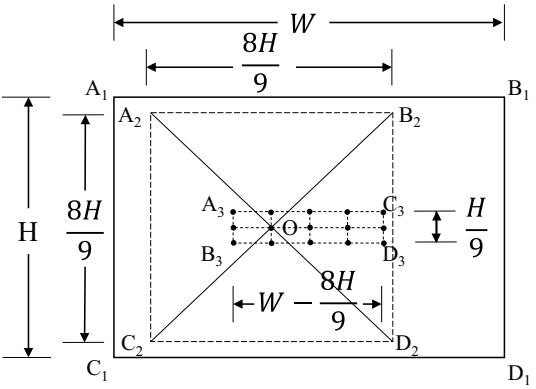

Fig. 4 The cropping scheme to balance DVS distribution. $A_1B_1C_1D_1$ is the original photo, and the $A_2B_2C_2D_2$ is the cropping region. The solid points, which divide the $A_3B_3C_3D_3$ into several same parts, represent where the O (center point of $A_2B_2C_2D_2$) could be. This figure demonstrates how to crop 15 images from the original photo with a DVS of six

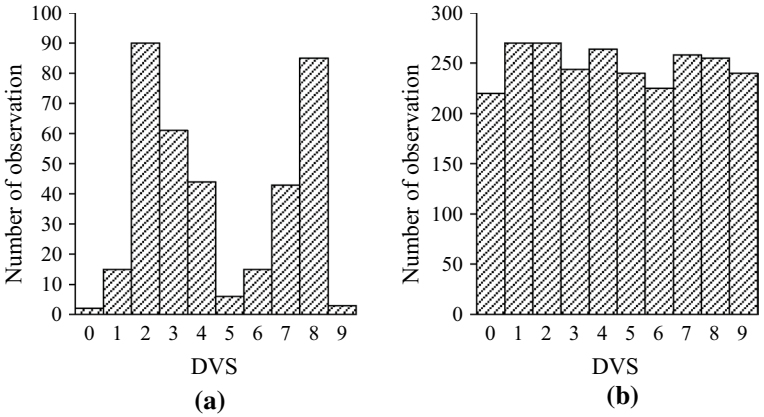

Fig. 5 Rice DVS distribution of training set: **a** DVS distribution of original statistic; **b** DVS distribution after cropping

The performance of the different models was evaluated in terms of overall accuracy (ACC, Eq. 1) and mean absolute error (MAE, Eq. 2):

$$\mathrm{ACC} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP}) \tag{1}$$

$$\mathrm{MAE} = \sum_{i=1}^{n} \hat{y} - y \tag{2}$$

where *TP* denotes the true positive, FP denotes the false positive, *n* denotes the total number of samples in the testing set, $\hat{y}$ denotes the predicted DVS, and *y* denotes the real DVS.

Furthermore, the ability to recognize a specific DVS was evaluated in terms of the F-score (F, Eq. (5)). A criterion is given that when the F-score of the DVS reached 0.75, the DVS was considered to have been correctly identified.

$$\mathrm{ACC_i} = \mathrm{TP_i}/\left(\mathrm{TP_i} + \mathrm{FP_i}\right) \tag{3}$$

$$\mathrm{Recall_i} = \mathrm{TP_i}/\left(\mathrm{TP_i} + \mathrm{FN_i}\right) \tag{4}$$

$$\mathrm{F_i} = 2 \times \mathrm{ACC_i} \times \mathrm{Recall_i}/\left(\mathrm{ACC_i} + \mathrm{Recall_i}\right) \tag{5}$$

Gcc-times series approach

The Green chromatic coordinate (Gcc, Eq. (6)), which can calculate the development of the canopy, is defined as the ratio of the green digital number (DN) to the sum of all digital numbers (or image brightness values) (Schwartz 2013):

$$\mathrm{Gcc} = \mathrm{Green\ DN}/(\mathrm{Red\ DN} + \mathrm{Green\ DN} + \mathrm{Blue\ DN}) \tag{6}$$

where DN is the constituent value of a given color in RGB color space.

Because the Gcc varied with both vertical and horizontal view angles, 12 time-series Gccs of a plot were derived from images shot at different angles, and the individual time-series Gcc were smoothed using a fitted locally weighted regression (LOWSS) model with a fixed span of 0.4 (Fig. 6a). The time-series Gcc was normalized by dividing by maximum value of the series to render the data comparable because the Gcc might have varied in the same DVS when crops were affected by planting density, fertilizer, and other factors.

A set of thresholds were derived to link the Gcc values with the DVS. For example, the smoothed and normalized time-series Gcc of the training set were compiled. A threshold was then used between the minimum and maximum values of the set. The DVS of the point located to the left of the maximum value of the time-series Gcc and below the threshold was assumed to be zero, and the DVS values of other points were assumed to be one. The thresholds for DVS = 0 and 1 were determined when the accuracy reached its highest value, and thresholds for the other DVS values were determined in the same way (Fig. 6b).

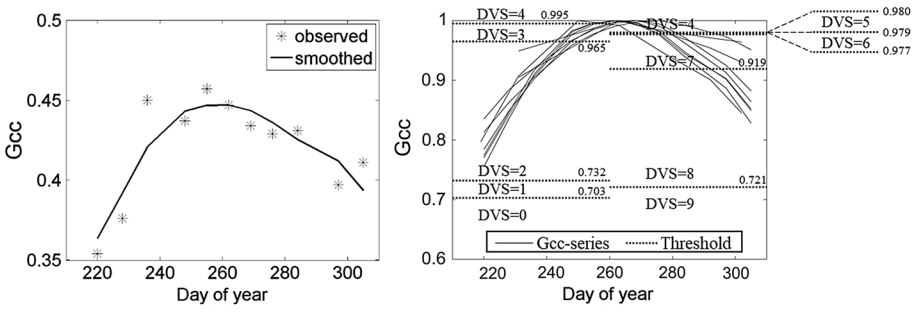

Fig. 6 The Gcc approach to determine development stage. **a** Smoothed and normalized Green Chromatic Coordinate (Gcc) time series; **b** The thresholds of 10 development stages and part of time-series-Gcc from the dataset with a vertical view angel of 20° and horizon view angle of 45°

Manually extracted feature combining SVM approach

Color and texture features are two types of manually extracted features used for image classification. The color features used here consisted of mean and variance values, and the texture features consisted of contrast, correlation, energy, and homogeneity as derived from a gray-level co-occurrence matrix (GLCM) at 0°, 45°, 90°, and 135°. These features were extracted from 15 channels in five color spaces (R, G, B in RGB color space, H, S, V in HSV color space, H, S, I in HSI color space, L, a*, b* from CIE L*a*b* color space, and Y, Cb, Cr from YCbCr color space). This resulted in a vector consisting of 270 feature values (2 features × 15 channels + 4 features × 4 directions × 15 channels).

To compare the conventional classifiers and the DCNN, these features were processed through a SVM to determine the DVS, and calculated the ACC and MAE of the manually extracted feature combining SVM approach (MF-SVM).

Deep convolutional neural network approach

To integrate images obtained at different view angles, an architecture was proposed that uses four separate branches to process four RGB images taken at four angles (A, B, C, D) (Fig. 7). A single branch is an AlexNet (Krizhevsky et al. 2012). Therefore, a pre-trained AlexNet can be fine-tuned to this dataset. A concatenation layer is designed to combine the features extracted from A, B, C, and D. Every branch provides a DVS prediction while the combined feature provides another based on the information of the four input images.

The parameters of the DCNN network were optimized using the back-propagation algorithm. A loss function and an optimizer were thus essential for network parameter optimization. Five cross-entropy losses were employed after the five predictions. In addition, three strategies— training from scratch (TFC), fine-tuning (FT), and two-step fine-tuning (TSFT)—were used to the optimize the network based on stochastic gradient descent with momentum (SGDM) acceleration

(Sutskever et al. 2013). The differences among these strategies in terms of data and the training process are described below.

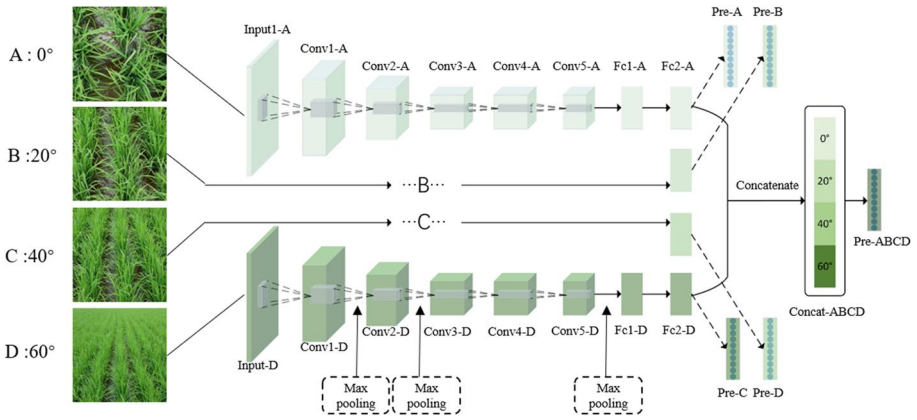

Fig. 7 The architecture of DCNN in this study. The DCNN is composed of four branches processing images from different view angles separately. The size of each layer is: Input—-256 × 256 × 3, Conv1-62 × 62 × 96, Conv2-30 × 30 × 256, Conv3-14 × 14 × 384, Conv4-14 × 14 × 384, Conv5-14 × 14 × 256, Fc1-4096, Fc2-4096 (Color figure online)

Training from scratch

Because the images in the datasets were captured at four vertical angles, they were divided into four groups according to angle. In the training process, the four groups were shuffled according to the same random number list and divided into mini-batches to feed to the networks. This means that the four images fed to the DCNN network were from the same observations at the same horizontal angle but different vertical angles.

Fine-tuning and two-step fine-tuning

Fine-tuning a network is based on the concept of transfer learning (Hope 2012). The general fine-tuning approach is to train a DCNN model with a classification function at the top of the network in a dataset with a large domain. Some layers are then replaced with new ones, the parameters of which are randomly initialized. Finally, a specific dataset with a small domain is employed to optimize the parameters of the network. In this study, the FT approach was first used to evaluate its capability to improve the performance of a pre-trained model. The parameters of four branches were all optimized by fine-tuning the Alexnet (BVLC AlexNet) (Jia et al. 2014) to the images of 0°, 20°, 40° and 60°, respectively. However, this strategy cannot make full use of all data because each branch only employed images at one view angle while the images at other view angles could also help to optimize the parameters. Thus, a two-step fine-tuning strategy was proposed to utilize the images of all view angles. The TSFT was divided into two steps.

First, parameters of each branch were optimized by fine-tuning the pre-trained AlexNet (BVLC AlexNet) to images at the other three view angles. For example, images at angles A (0°), B (20°), and C (40°) were mixed, shuffled, and fed into the pre-trained AlexNet, resulting a pre-fine-tuning AlexNet dedicated to branch D. This step was repeated four times to get four pre-fine-tuning networks for the four branches for next step.

Second, the parameter of the four pre-fine-tuning AlexNets were used to initialize each branch of the network while the parameters of the pre-ABCD layer were randomly initialized. The four images were then fed to the network in the same way as in TFS.

Testing

In the testing phase, the prediction was the average result of sample crops from the original image. For example, the four original images with the same horizontal angle were cropped six times, resulting in 24 samples. Four samples from four original images were fed into the network, and this procedure was repeated six times to get six predictions that were averaged and rounded to the nearest whole number as the final prediction.

Results

Classification of development stages using time-series Gcc

The results for different view angles using the time-series Gcc approach are presented in Table 2 in terms of ACC and MAE. Overall, the time-series Gcc recorded an acceptable

Table 2 The ACC and MAE of time-series-Gcc approach

ACC/MAE	Horizon view angle		
	a: 0°	b: 45°	c: 90°
Vertical view angle			
A: 0°	0.722/0.356	0.726/0.358	0.701/0.395
B: 20°	0.702/0.377	0.737/0.317	0.695/0.366
C: 40°	0.695/0.390	0.713/0.343	0.721/0.340
D: 60°	0.660/0.450	0.660/0.452	0.652/0.495

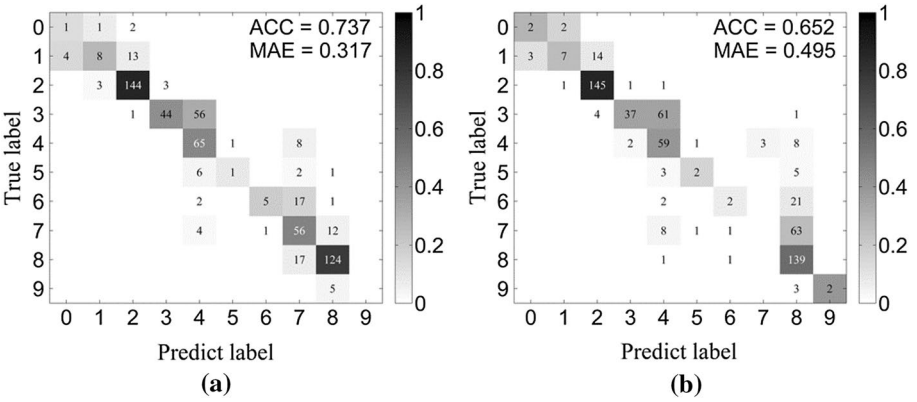

Fig. 8 The best and worst performance of time-series-Gcc approach. **a** Confusion matrix yielded by the Bb set; **b** confusion matrix yielded by the Dc set

ability to identify DVS. Of the 12 view angles, angle Bb produced the best result with the highest ACC (0.737) (Fig. 8a) and the smallest MAE (0.317). Even the worst result delivered a high ACC of 0.652 (Fig. 8b) and an MAE of 0.495.

However, the time-series Gcc approach classifies the DVS according to the threshold of greenness, which is sensitive to crop growth and senescence. According to Table 3, only the tillering stage (DVS = 2) and ripening stage (DVS = 8) were correctly identified, with F-scores of 0.93 and 0.87, respectively. This is because greenness increased rapidly in the tillering stage and decreased abruptly in the ripening stage, whereas only a small change occurred in the other stages. This means that the time-series Gcc can only be used to monitor the DVS when greenness changes rapidly.

Two further weaknesses rendered the time-series Gcc unsuitable for DVS monitoring. First, the requirement of the time series limits its agricultural application. Second, the time-series Gcc approach uses time-series data to fit a curve, where these data need to be normalized by dividing by the maximum value of the series. The missing data, especially a missing maximum value, introduces error to the fitting curve.

Classification of development stages by MF-SVM approach

The results obtained by the MF-SVM approach based on 270 features were better than those of the time-series Gcc approach (Table 4), and angle C yielded best results with an

Table 3 The F-score of different view angles and different approaches

F-score		DVS									
Approach	Angle	0	1	2	3	4	5	6	7	8	9
		Germination	Leaf development	Tillering	Stem elongation	Booting	heading	Flowering	Development of fruit	Ripening	Senescence
Gcc	Bb	0.22	0.43	0.93	0.59	0.63	0.17	0.32	0.65	0.87	0.00
MF-SVM	C	0.80	0.55	0.82	0.88	0.85	0.50	0.76	0.78	0.91	0.50
DCNN	ABCD(L)	0.00	0.71	0.94	0.81	0.61	0.14	0.40	0.89	0.96	0.80
DCNN	ABCD(M)	0.80	0.76	0.96	0.92	0.84	0.00	0.67	0.95	0.97	0.80
DCNN	ABCD(H)	1.00	0.81	0.97	0.93	0.85	0.25	0.86	0.93	0.96	0.50

L, M and H represent low, middle and high resolution

ACC of 0.817 and a MAE of 0.208. Furthermore, the F-scores of angle C listed in Table 3 indicate that seven stages of rice could be identified—germination, tillering, stem elongation, booting, flowering, development of fruit, and ripening.

Three feature vectors consisting of different numbers of features were used to investigate the effects of the number of features on DVS detection. According to Fig. 9, angle C of the MF-SVM, which delivered the best performance when the number of features was 270, yielded an ACC ranging from 0.470 to 0.735 when the number of channels varied from one to six. This is because the features extracted from the GLCM contained phenological information that helps SVM to classify the DVS.

In general, the MF-SVM is a better choice than the Gcc approach for the following two advantages. First, it eliminates time-series data, and only a single photo can be used for DVS identification, which makes it possible to obtain phenological information in real time. Second, if more features can be designed and extracted from photos, the results of the classification can be rendered more accurate.

Classification of development stages using deep convolutional network

Results of DCNN based on middle-resolution dataset

The DCNN and MF-SVM are similar as both use a classifier to classify the DVS based on features extracted from the image. The difference is that the SVM uses manually extracted features while the DCNN uses features automatically extracted by the network. Therefore, the performance of the DCNN depends on whether the network parameters are well optimized, which is related to the training strategy. The results of the DCNN using the middle-resolution dataset are presented in Fig. 10, which shows that its performance was inferior to that of the MF-SVM if the model was trained from scratch. Although combining the features of the four branches, the best result was not remarkable (with ACC = 0.799 and MAE = 0.263), and the results of the four single branches were all worse than that of the MF-SVM. However, after the TSFT strategy was employed, each branch delivered outstanding performance, with ACC values ranging from 0.829 to 0.857 and MAE ranging from 0.245 to 0.16. The results of the four single branches indicated that a well-trained DCNN model can better extract features to classify the DVS than the MF-SVM approach. Moreover, the performance of the DCNN significantly improved after combining features from the four branches, with ACC = 0.901 and MAE = 0.122.

Furthermore, the DCNN improved the F-scores of some stages (Table 3), which made it possible to classify more development stages correctly. In addition to the seven stages detected by the MF-SVM, two stages including leaf development stage and senescence stage were correctly identified, the F-scores of which improved from 0.55 to 0.76 and 0.50 to 0.80, respectively. However, the result of the flowering stage deteriorated slightly,

Table 4 The ACC and MAE of MF-SVM approach

	Vertical view angle			
	A: 0°	B: 20°	C: 40°	D: 60°
ACC	0.749	0.764	0.817	0.790
MAE	0.314	0.266	0.208	0.246

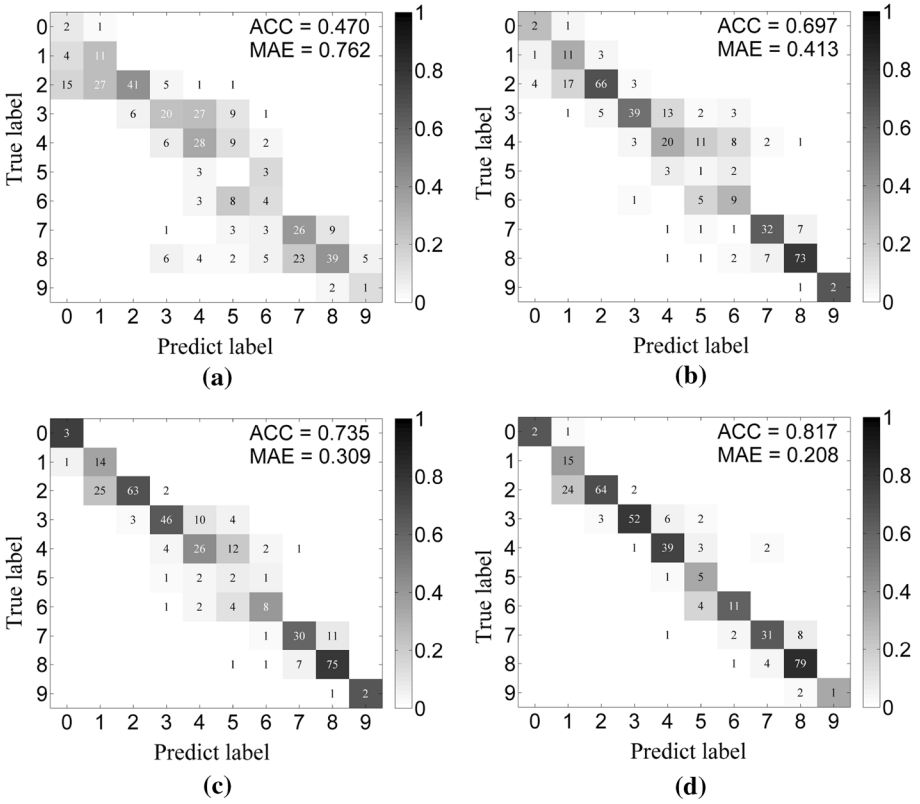

Fig. 9 The performance of GLCM based on different numbers of channel and feature. **a–d** are results yielded by18 features from one channel, 54 features from 3 channels, 108 features from 6 channels, and 270 features from 15 channels, respectively

making this stage difficult to identify. Thus, eight of 10 DVS were identified when four angle photos were combined.

Images at different spatial resolutions for DVS classification

Identifying development stages is particularly difficult during transitions from stage to another because the looming features are too small to distinguish at the beginning of a stage. Improving image resolution is beneficial for DVS detection. As shown in Fig. 11, the dataset with low-resolution images yielded the worst result (with ACC = 0.825 and MAE = 0.198), whereas high-resolution images improved the results slightly (with ACC = 0.912 and MAE = 0.102) compared with middle-resolution images. The results show that a higher image resolution remarkably improved classification ability, especially for stages with small size features. For instance, it was difficult to extract features from images of the booting stage and flowering stage because the former's feature (the flag leaf) is difficult to distinguish from other leaves when it is small, and the latter is characterized by anthers that are very small. After improving the image resolution from low to high, the F-scores of the booting and flowering stages increased by 0.24 and 0.46.

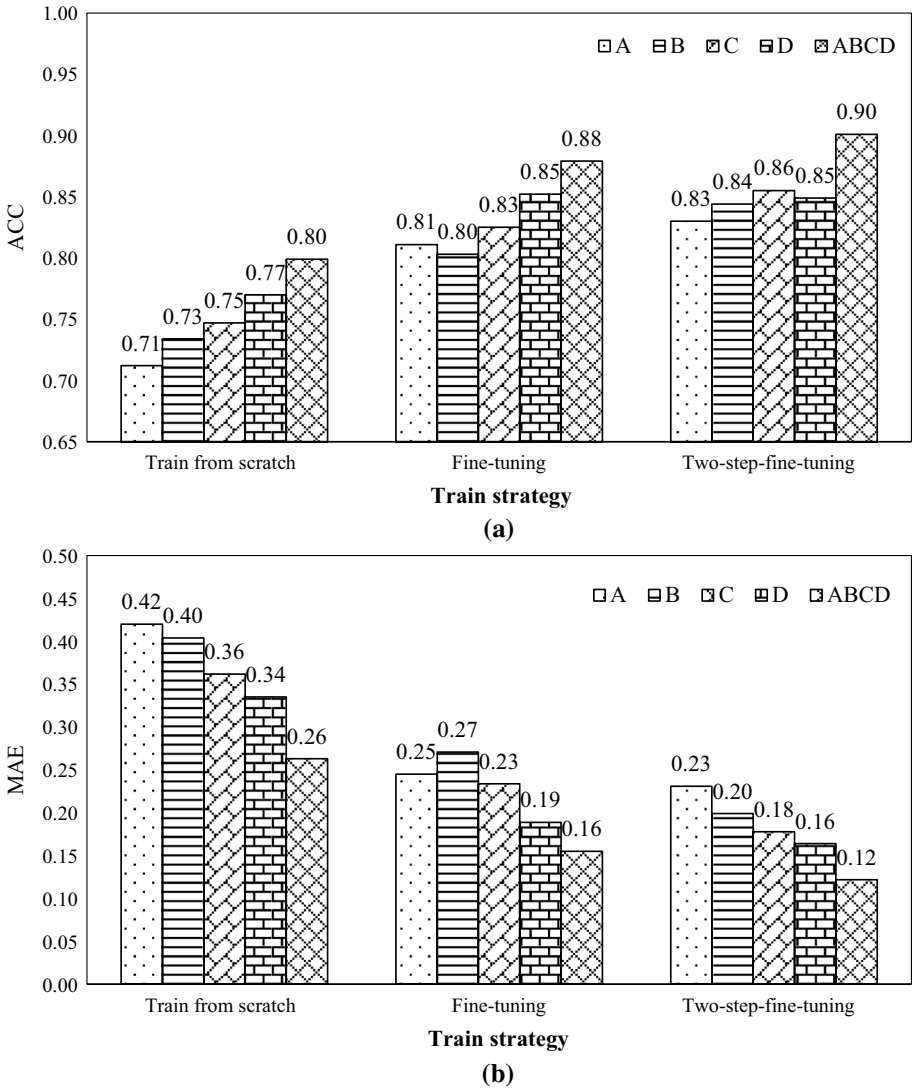

Fig. 10 ACC and MAE of different train strategy and different branch. A, B, C, D represent the 0°, 20°, 40° and 60°. ABCD represents the branch concentrates the features derived from four angles

However, the heading stage was still difficult to identify although its F-score increased a little. As shown in Fig. 11d, the stage developed quickly between BBCH40 and BBCH60, and thus some clusters were still in the booting stage while others had transitioned to the heading stage or even the flowering stage. Furthermore, according to Fig. 5, the number of images of the heading stage was too small to train the network. Bai et al. (2018) have proposed a method to identify the heading stage by cropping the high-resolution photos into mini-patches that are fed into an SVM and a DCNN to identify them as spike or non-spike, and the heading stage is identified by the number of mini-patches considered to be spikes. This approach is more like one based on manually extracted features than an end-to-end

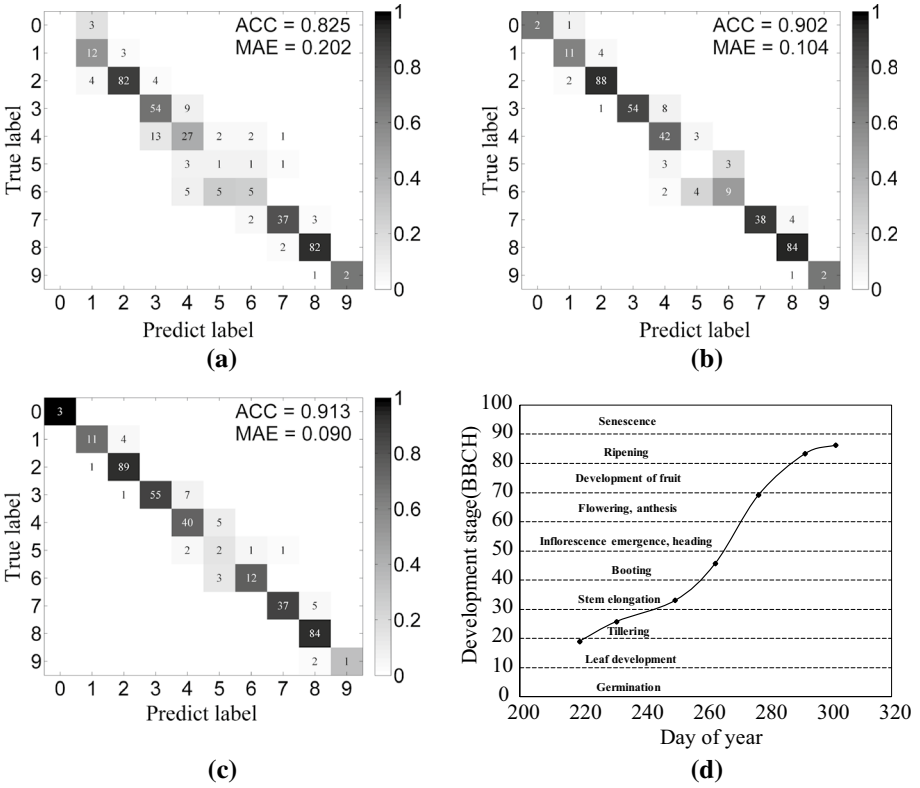

Fig. 11 The confusion matrix yielded by three datasets with low (**a**), middle (**b**), and high (**c**) images. The three results were all yielded by TSFT-network. **d** Picture of the average development stage of experiment plots changing over time

method based on the DCNN. This is a good means of identifying stages with features of small size. Therefore, future work should focus on integrating manually extracted features with the DCNN to improve the classification of development stages.

Discussion

Results of different training strategies

Because the performance of three strategies varied greatly, it is worthwhile comparing the differences among these strategies. The discussion in this section and "Classification of development stages by MF-SVM approach" section is based on the results obtained on the medium-resolution image dataset.

According to Fig. 10, the results were poor when the dataset was used to train the network directly, as the highest ACC of the single branch was only 0.77. Tajbakhsh et al. (2016) have noted that training a DCNN from scratch is difficult. Because it requires a large number of labeled training data (e.g., the DVS), which is difficult in agriculture because of the scarcity of labeled images. Thus, unless large datasets containing millions

of labeled data are available, training from scratch is not a good way to optimize the model parameters.

To overcome the shortcomings of training from scratch, the pre-trained Alexnet (BVLC AlexNet) was used. This greatly improved classification accuracy, with the ACC of each branch ranging from 0.803 to 0.852 and MAE from 0.271 to 0.189. Although the pre-trained model can help networks converge quickly and improve the performance of the model, this general fine-tuning procedure did not make full use of all data because only a quarter of photos were used to fine-tune the parameters of a single branch.

The TSFT strategy was used to solve the above problem. Recent research by Azizpour et al. (2015) suggests that the success of knowledge transfer depends on the distance, or dissimilarity, between the database on which a DCNN is trained and that to which the knowledge is to be transferred. Compared with the distance between a natural object in ImageNet and photos taken at a view angle, the distance between photos taken at different angles was smaller. Thus, the two-step strategy, the first step of which is to adapt the pre-trained model to the domain of paddy rice by fine-tuning parameters on the three of four photos and second is to fine-tune the parameters to fit the specific view angle, further improved the capability for DVS classification by reducing the MAE of four branches by 5.7%, 26.6%, 23.9%, and 13.2%, respectively.

Results at different view angles

Owing to limited research on the impact of methods of field data acquisition on DVS estimation, this dataset containing photos shot from four vertical angles can provide guidance for collecting high-value field data in future research.

Figure 10 demonstrates that the results of angle D were always better than those of the other three angles while angle A delivered the worst results. Note that regardless of the training strategy used, the performance improved as the view angle increased. Two factors might be related to this phenomenon. First, the photos were more representative as the view angle increased. The development stages of different rice crops were varied due to the heterogeneity of water and fertilizer distribution. As Fig. 12 shows, more rice clusters were captured in images with angle increasing. This reduced uncertainty caused by heterogeneity. Second, photos shot at large angles contained more features at different scales, which provided more information to the network to estimate phenology. Figure 12 shows that photos shot at 0° provided information only at the scale of organ (e.g., type and color), and those shot at 20° and 40° provided information at a median scale (e.g., proportion of rice ear, stem, and leaf, and the degree of bending of the rice ear) in addition to that at the organ scale. Photos shot at 60° provided organ information at the close range, plant information at the middle range, and group information (e.g., canopy greenness and closure level) at the global range. Note that the best result was obtained by photos shot at 60° because once the angle exceeded 60°, there would be a high probability for images to capture extraneous objects irrelevant to DVS. Thus, the best view angle at a height of 1.5 m for small plots was 60°.

Although photos taken at large angles usually delivered better results than those at small angles, a counterexample is shown in Fig. 10 where angle B gave the worst result when the FT strategy was used. However, Fig. 10 shows that when the TSFT strategy was used, branch A was suppressed because “knowledge” gleaned from photos at the other three angles using branch B was more valuable than that obtained from branch A. As shown in Fig. 13, the view angle of Fig. 12a was larger than that of Fig. 13b, but both belong to the

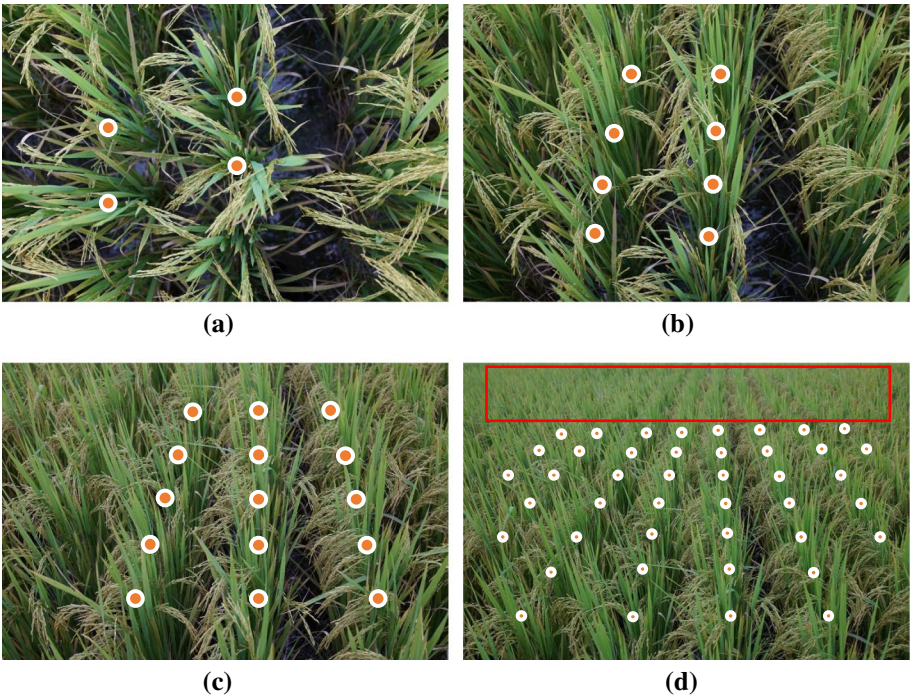

Fig. 12 Comparison of photos shot at different vertical angles. **a**–**d** were the photos shot at 0°, 20°, 40°, 60°. One point in the photo represents one rice cluster while 4, 8, 15, 42 clusters are contained in these four photos. The region surrounded by red rectangle provides information of canopy greenness or canopy closure level (Color figure online)

20° dataset. This problem arose because the handheld camera could not precisely control the view angle, especially at 20° and 40° because 0° could be determined by gravity and 60° according to the horizon. Figure 10 shows that after learning from photos taken at the other three angles, the MAE of branches B and C decreased by 0.7 and 0.5, respectively, while that of A and D decreased by only 0.3 and 0.2. Therefore, if the view angle can be controlled when collecting training data, the performance of the network can be improved.

Robustness of well-trained model against view angle

Error in the estimated view angle did not arise only in the training data, but also in data used to identify stages of development. To evaluate the robustness of the network against random view angle, images taken at incorrect angles were fed into the network. For example, branches A, B, C, and D were fed images at angles ABCD, BCDA, CDAB, and DABC, respectively.

The results in Fig. 14a and c show that when the model was trained by FT strategy, its performance deteriorated if the angle of the image did not match the branch, and worsened further as the difference between the angles increased. For example, the ACC values obtained by branch A were 0.81, 0.78, 0.59, and 0.44 for images A, B, C, and D, respectively. However, with TSFT strategy, the robustness of the network improved significantly. As shown in Fig. 14b and d, performance gradually improved with an

(a) (b)

Fig. 13 The uncertainty of view angle caused by hand. **a**, **b** were both shot with a vertical angle of 20°

increase in the angle of images from A to D, but the results did not change considerably across branches. This means that the performance of the model no longer depended on the matching level between the input image and the branch, but on the information contained in the input image itself. Therefore, the TSFT strategy improves the performance of each branch and makes the network more robust against the random angle at which the image is taken.

Features extracted by network

The artificial neural network (ANN) is often criticized for its lack of interpretability. Therefore, the gradient-weighted class activation mapping (Grad-CAM) (Selvaraju et al. 2017) was used to explain how DCNN identified the development stages of rice from images.

As shown in the input images in Fig. 15, the rice crop in each DVS has distinctive characteristics, such as yellow straws after harvest for DVS = 0, small rice clusters and a water surface for DVS = 1, and large rice cluster and the bare ground for DVS = 2. The class activation mapping (CAM) in Fig. 15 indicates that the DCNN correctly found and extracted these features as basis for phenology classification. However, it was difficult to understand the difference between the red area in images with DVS = 3 and 4, although the DCNN model correctly classified these two stages based on features contained in these areas. For DVS = 5, 6, 7, 8, and 9, the features were mainly concentrated in spikes, with only small differences between spikes in the different stage. Thus, it was easy for the DCNN to focus on the spikes but difficult to identify the small differences that are useful in distinguishing similar stages. However, the DCNN model could classify images with DVS = 7 and 8 although their features appeared similar, while the images with DVS = 5, 6, and 9 were difficult to identify even if the differences among them were more prominent. This might have occurred due to the difference in the numbers of samples in the training set. According to Fig. 16, the F-score improved as the number of samples increased. It was difficult for the network to acquire a enough knowledge from the small dataset, where the numbers of training samples with DVS = 5, 6, and 9 were only 24, 60, and 12 respectively. The datasets for DVS = 7 and 8 contained 168 and 336 photos, respectively. Therefore, when the features were not prominent enough for the network to learn, increasing the number of samples can compensate for this deficiency.

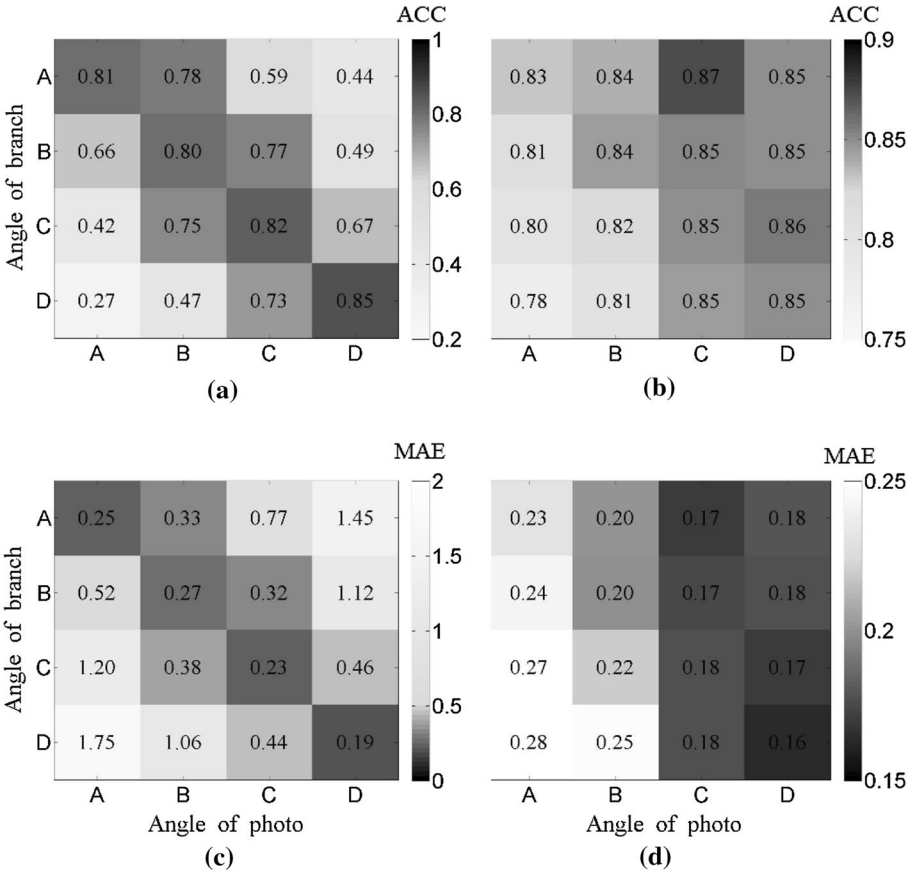

Fig. 14 The ACC and MAE for different branches to identify development stage based on images of different angles. **a**, **c** are obtained by the FT-network. **b**, **d** are obtained by the TSFT-network

Conclusion

This study proposed an approach for the identification of development stages of rice based on handheld camera RGB photos. To the best of our knowledge, this is the first investigation of DVS classification by using random view angle RGB photos taken by a handheld camera. The proposed DCNN structure consists of four separate branches to process RGB images shot from four vertical angles. Compared with the time-series Gcc method (ACC = 0.731, MAE = 0.317) and the MF-SVM method (ACC = 0.817, MAE = 0.208), the DCNN method classified the DVS more accurately (ACC = 0.913) and with a smaller error (MAE = 0.090). It can thus be used by smallholder farmers to identify phenology using handheld smartphones in real time.

Furthermore, images taken at different view angles, different model training strategies, and interpretations of predictions of DCNN model were investigated. The results showed that photos taken at a large angle were more valuable because they contained more information than photos taken at a small angle. The proposed two-step-fine-tuning strategy greatly improved the robustness of the model and lessened the influence of the uncertainty

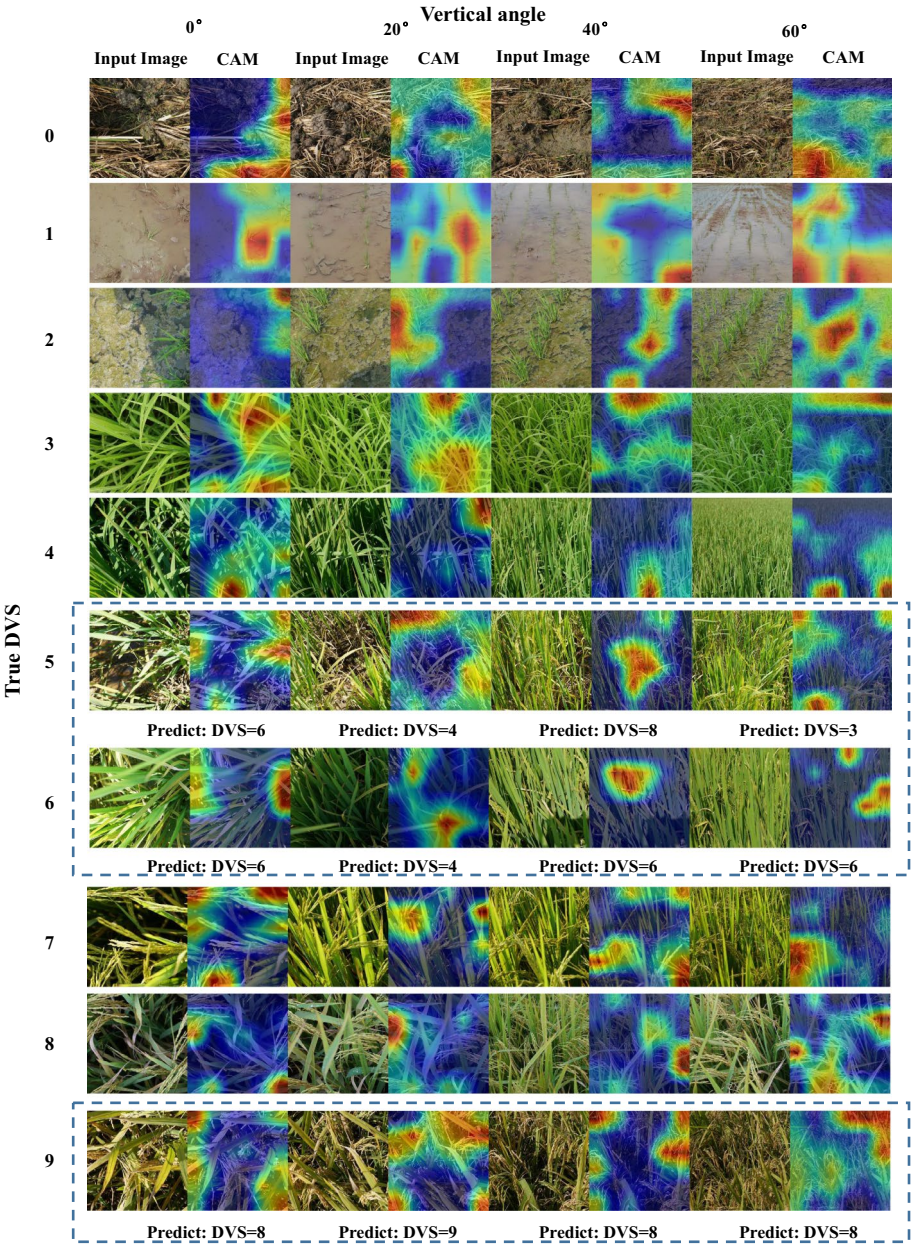

Fig. 15 The Grad-CAM derived from the high-resolution dataset, images from left to right come from different angles in one plot. The development stages of these images are all identified correctly except that of images surrounded by the blue dotted frames. The true DVS of these images are shown on the left and predictions of image surrounded by the blue dotted frames are shown under images and the Grad-CAMs. Color in the CAM represents the importance of the region for DCNN and the red area contributes a lot to the final classification result while the contribution of the blue area is small (Color figure online)

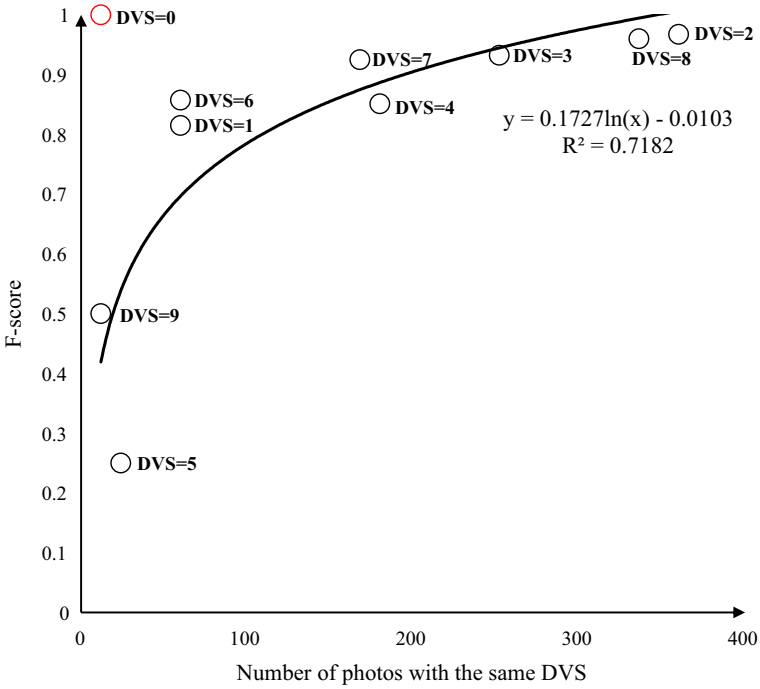

Fig. 16 The relation between F-score yielded by high-resolution images and the number of original photos in the training set. The curve is a fitted line of nine black hollow points except for the red hollow point (the easy-distinguished features in images of DVS = 0, such as bare soil and yellow stalks, make the F-score high even with a small number of images) (Color figure online)

of view angle. Grad-CAM showed that the network can automatically find information related to development stage from images. This study offers a promising deep learning approach for the real-time identification of development stages of rice on small plots as RGB photos at a high spatial resolution become increasingly available.

Acknowledgements This study was supported by the National Natural Science Foundation of China Grant 51861125202 and 51629901, and the Key Research and Development Program in Guangxi Grant AB16380257.

References

Azizpour, H., Razavian, A. S., Sullivan, J., Maki, A., & Carlsson, S. (2015). From generic to specific deep representations for visual recognition. In *IEEE Conference on computer vision and pattern recognition workshops*, October 2015 (pp. 36–45). https://doi.org/10.1109/CVPRW.2015.7301270.

Bai, X., Cao, Z., Zhao, L., Zhang, J., Lv, C., Li, C., & Xie, J. (2018). Rice heading stage automatic observation by multi-classifier cascade based rice spike detection method. *Agricultural and Forest Meteorology, 259*, 21360–270. https://doi.org/10.1016/j.agrformet.2018.05.001

Chen, S. W., Shivakumar, S. S., Dcunha, S., Das, J., Okon, E., Qu, C., et al. (2017). Counting apples and oranges with deep learning: A data-driven approach. *IEEE Robotics and Automation Letters, 2,* 781–788. https://doi.org/10.1109/LRA.2017.2651944

Cui, Z., Zhang, H., Chen, X., Zhang, C., Ma, W., Huang, C., et al. (2018). Pursuing sustainable productivity with millions of smallholder farmers. *Nature, 555,* 363–366. https://doi.org/10.1038/nature25785

Dyson, J., Mancini, A., Frontoni, E., & Zingaretti, P. (2019). Deep learning for soil and crop segmentation from remotely sensed data. *Remote Sensing, 11,* 1859. https://doi.org/10.3390/rs11161859

Ferentinos, K. P. (2018). Deep learning models for plant disease detection and diagnosis. *Computers and Electronics in Agriculture, 145,* 311–318. https://doi.org/10.1016/j.compag.2018.01.009

Gebbers, R., & Adamchuk, V. I. (2010). Precision agriculture and food security. *Science*. https://doi.org/10.1126/science.1183899

Guo, W., Fukatsu, T., & Ninomiya, S. (2015). Automated characterization of flowering dynamics in rice using field-acquired time-series RGB images. *Plant Methods, 11,* 7. https://doi.org/10.1186/s13007-015-0047-9

Hope, V. M. (2012). Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML-2011 workshop on unsupervised and transfer learning* (pp. 13–36). https://doi.org/10.1007/978-3-642-36657-4_1.

Huete, A., Miura, T., Yoshioka, H., Ratana, P., & Broich, M. (2013). Indices of vegetation activity. In J. M. Hanes (Ed.), *Biophysical applications of satellite remote sensing* (pp. 1–41). Berlin: Springer. https://doi.org/10.1007/978-3-642-25047-7_1.

Hufkens, K., Melaas, E. K., Foster, T., Robles, M., Mann, M. L., Kramer, B., & Ceballos, F. (2018). Monitoring crop phenology using a smartphone based near-surface remote sensing approach. *Agricultural and Forest Meteorology, 265,* 327–337. https://doi.org/10.1016/j.agrformet.2018.11.002

Ichardson, A. N. D. R., Raswell, B. O. H. B., Ollinger, D. A. Y. H., & Enkins, J. U. P. J. (2009). Near-surface remote sensing of spatial and temporal variation. *Ecological Applications, 19,* 1417–1428.

Jamieson, P. D., Brooking, I. R., Semenov, M. A., McMaster, G. S., White, J. W., & Porter, J. R. (2007). Reconciling alternative models of phenological development in winter wheat. *Field Crops Research , 103*(1), 36–41. https://doi.org/10.1016/j.fcr.2007.04.009

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., & Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. In *ACM conference on multimedia* (pp. 675-678). https://doi.org/10.1145/2647868.2654889.

Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture, 147,* 70–90. https://doi.org/10.1016/j.compag.2018.02.016

Klosterman, S., Melaas, E., Wang, J., Martinez, A., Frederick, S., O'Keefe, J., et al. (2018). Fine-scale perspectives on landscape phenology from unmanned aerial vehicle (UAV) photography. *Agricultural and Forest Meteorology, 248,* 397–407. https://doi.org/10.1016/j.agrformet.2017.10.015

Koirala, A., Walsh, K. B., Wang, Z., & McCarthy, C. (2019). Deep learning for real-time fruit detection and orchard fruit load estimation: Benchmarking of 'MangoYOLO.' Precision Agriculture, 20, 1107–1135. https://doi.org/10.1007/s11119-019-09642-0.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Handbook of approximation algorithms and metaheuristics* (pp. 60-1–60-16). London: Chapman & Hall. https://doi.org/10.1201/9781420010749.

Liu, X., Chen, S. W., Aditya, S., Sivakumar, N., Dcunha, S., Qu, C., Taylor, C. J., Das, J., & Kumar, V. (2018). Robust fruit counting: Combining deep learning, tracking, and structure from motion. In *IEEE international conference on intelligent robots and systems* (pp. 1045–1052). https://doi.org/10.1109/IROS.2018.8594239.

Liu, X., Li, S., Kan, M., Zhang, J., Wu, S., Liu, W., Han, H., Shan, S., & Chen, X. (2015). AgeNet: Deeply learned regressor and classifier for robust apparent age estimation. In *Proceedings of IEEE internationalconference on computer vision workshops* (pp. 16–24). https://doi.org/10.1109/ICCVW.2015.42.

Lowder, S. K., Skoet, J., & Raney, T. (2016). The number, size, and distribution of farms, smallholder farms, and family farms worldwide. *World Development, 87,* 16–29. https://doi.org/10.1016/j.worlddev.2015.10.041

Ma, J., Li, Y., Chen, Y., Du, K., Zheng, F., Zhang, L., & Sun, Z. (2019). Estimating above ground biomass of winter wheat at early growth stages using digital images and deep convolutional neural network. *European Journal of Agronomy, 103*, 117–129. https://doi.org/10.1016/j.eja.2018.12.004.

Madec, S., Jin, X., Lu, H., De Solan, B., Liu, S., Duyme, F., et al. (2019). Ear density estimation from high resolution RGB imagery using deep learning technique. *Agricultural and Forest Meteorology, 264*, 225–234. https://doi.org/10.1016/j.agrformet.2018.10.013

Milioto, A., Lottes, P., & Stachniss, C. (2018). Real-Time semantic segmentation of crop and weed for precision agriculture robots leveraging background knowledge in CNNs. In *Proceedings of IEEE international conference on robotics and automation* (pp. 2229–2235). https://doi.org/10.1109/ICRA.2018.8460962.

Park, J. Y., Muller-Landau, H. C., Lichstein, J. W., Rifai, S. W., Dandois, J. P., & Bohlman, S. A. (2019). Quantifying leaf phenology of individual trees and species in a tropical forest using unmanned aerial vehicle (UAV) images. *Remote Sensing, 11,* 1534. https://doi.org/10.3390/rs11131534

Patrício, D. I., & Rieder, R. (2018). Computer vision and artificial intelligence in precision agriculture for grain crops: A systematic review. *Computers and Electronics in Agriculture, 153,* 69–81. https://doi.org/10.1016/j.compag.2018.08.001

Perez, L., & Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. arXiv preprint. arXiv:1712.04621.

Petach, A. R., Toomey, M., Aubrecht, D. M., & Richardson, A. D. (2014). Monitoring vegetation phenology using an infrared-enabled security camera. *Agricultural and Forest Meteorology, 195–196*, 143–151. https://doi.org/10.1016/j.agrformet.2014.05.008

Putra, B. T. W., & Soni, P. (2019). Improving nitrogen assessment with an RGB camera across uncertain natural light from above-canopy measurements. *Precision Agriculture, 13*(3), 285–301. https://doi.org/10.1007/s11119-019-09656-8

Ricciardi, V., Ramankutty, N., Mehrabi, Z., & Jarvis, L. (2018). How much of the world's food do smallholders produce? *Global Food Security, 17,* 64–72. https://doi.org/10.1016/j.gfs.2018.05.002

Sa, I., Popović, M., Khanna, R., Chen, Z., Lottes, P., Liebisch, F., et al. (2018). WeedMap: A large-scale semantic weed mapping framework using aerial multispectral imaging and deep neural network for precision farming. *Remote Sensing, 10*(9), 1423. https://doi.org/10.3390/rs10091423

Sakamoto, T., Gitelson, A. A., Wardlow, B. D., Arkebauer, T. J., Verma, S. B., Suyker, A. E., & Shibayama, M. (2012). Application of day and night digital photographs for estimating maize biophysical characteristics. *Precision Agriculture, 13*(3), 285–301. https://doi.org/10.1007/s11119-011-9246-1

Schwartz, M. D. (Ed.). (2013). *Phenology: An integrative environmental science* (pp. 548–550). Dordrecht: Kluwer.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, October 2017 (pp. 618–626). https://doi.org/10.1109/ICCV.2017.74.

Sladojevic, S., Arsenovic, M., Anderla, A., Culibrk, D., & Stefanovic, D. (2016). Deep neural networks based recognition of plant diseases by leaf image classification. *Computational Intelligence and Neuroscience, 2016*, Article 3289801. https://doi.org/10.1155/2016/3289801.

Smith, A. R. (1978). Color gamut transform pairs. *ACM SIGGRAPH Computer Graphics, 12*, 12–19. https://doi.org/10.1145/965139.807361

Sonnentag, O., Hufkens, K., Teshera-Sterne, C., Young, A. M., Friedl, M., Braswell, B. H., et al. (2012). Digital repeat photography for phenological research in forest ecosystems. *Agricultural and Forest Meteorology, 152*, 159–177. https://doi.org/10.1016/j.agrformet.2011.09.009

Lancashire, P. D., Bleiholder, H., Boom, T. V. D., Langelüddeke, P., Stauss, R., WEBER, E., & Witzenberger, A. (1991). A uniform decimal code for growth stages of crops and weeds. *Annals of Applied Biology, 119*(3), 561–601. https://doi.org/10.1111/j.1744-7348.1991.tb04895.x

Stein, M., Bargoti, S., & Underwood, J. (2016). Image based mango fruit detection, localisation and yield estimation using multiple view geometry. *Sensors (Switzerland), 16*, 1915. https://doi.org/10.3390/s16111915

Sunoj, S., Igathinathane, C., & Hendrickson, J. (2016). Monitoring plant phenology using phenocam: A review. In *ASABE annual international meeting* (pp. 1–9). https://doi.org/10.13031/aim.20162461829.

Sutskever, I., Martens, J., Dahl, G., & Hinton, G. (2013). On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th international conference on machine learning* (PMLR) (Vol. 28(3), pp. 1139–1147). https://doi.org/10.1017/CBO9781316423936.

Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., & Liang, J. (2016). Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Transactions on Medical Imaging, 35*, 1299–1312. https://doi.org/10.1109/TMI.2016.2535302

White, M. A., de Beurs, K. M., Didan, K., Inouye, D. W., Richardson, A. D., Jensen, O. P., et al. (2009). Intercomparison, interpretation, and assessment of spring phenology in North America estimated from remote sensing for 1982–2006. *Global Change in Biology, 15,* 2335–2359. https://doi.org/10.1111/j.1365-2486.2009.01910.x

Yalcin, H. (2017). Plant phenology recognition using deep learning: Deep-pheno. In *2017 The sixth international conference on agro-geoinformatics* (pp. 1–5). https://doi.org/10.1109/Agro-Geoinformatics.2017.8046996.

Yang, Q., Shi, L., Han, J., Zha, Y., & Zhu, P. (2019). Deep convolutional neural networks for rice grain yield estimation at the ripening stage using UAV-based remotely sensed images. *Field Crops Research, 235*, 142–153. https://doi.org/10.1016/j.fcr.2019.02.022

Yoshida, S. (1981). Fundamentals of rice crop science. In *Growth and development of the rice plant*. Los Baños: International Rice Research Institute.

Yu, Z., Cao, Z., Wu, X., Bai, X., Qin, Y., Zhuo, W., et al. (2013). Automatic image-based detection technology for two critical growth stages of maize: Emergence and three-leaf stage. *Agricultural and Forest Meteorology, 174–175*, 65–84. https://doi.org/10.1016/j.agrformet.2013.02.011

Zhang, N., Wang, M., & Wang, N. (2002). Precision agriculture—a worldwide overview. *Computers and Electronics in Agriculture, 36*, 113–132. https://doi.org/10.1016/S0168-1699(02)00096-0.

Zhang, X., Friedl, M. A., & Schaaf, C. B. (2006). Global vegetation phenology from Moderate Resolution Imaging Spectroradiometer (MODIS): Evaluation of global patterns and comparison with in situ measurements. *Journal of Geophysical Research: Biogeosciences, 111*, 1–14. https://doi.org/10.1029/2006JG000217

Zhang, X., Jayavelu, S., Liu, L., Friedl, M. A., Henebry, G. M., Liu, Y., et al. (2018). Evaluation of land surface phenology from VIIRS data using time series of PhenoCam imagery. *Agricultural and Forest Meteorology, 256–257*, 137–149. https://doi.org/10.1016/j.agrformet.2018.03.003

Zheng, H., Cheng, T., Yao, X., Deng, X., Tian, Y., Cao, W., & Zhu, Y. (2016). Detection of rice phenology through time series analysis of ground-based spectral index data. *Field Crops Research, 198*, 131–139. https://doi.org/10.1016/j.fcr.2016.08.027

Zhu, Y., Cao, Z., Lu, H., Li, Y., & Xiao, Y. (2016). In-field automatic observation of wheat heading stage using computer vision. *Biosystems Engineering, 143*, 28–41. https://doi.org/10.1016/j.biosystemseng.2015.12.015