Proceedings of Engineering and Technology Innovation, vol. 26, 2024, pp. 01-17

Recognition of Ginger Seed Growth Stages Using a Two-Stage Deep Learning Approach

Yin-Syuen Tong[1,2], Tou-Hong Lee[2], Kin-Sam Yen[1,*]

[1]School of Mechanical Engineering, Universiti Sains Malaysia, Nibong Tebal, Malaysia

[2]TMS LITE Sendirian Berhad, Sungai Ara, Malaysia

Received 07 August 2023; received in revised form 04 October 2023; accepted 05 October 2023

DOI: https://doi.org/10.46604/peti.2023.12701

Abstract

Monitoring the growth of ginger seed relies on human experts due to the lack of salient features for effective recognition. In this study, a region-based convolutional neural network (R-CNN) hybrid detector-classifier model is developed to address the natural variations in ginger sprouts, enabling automatic recognition into three growth stages. Out of 1,746 images containing 2,277 sprout instances, the model predictions revealed significant confusion between growth stages, aligning with the human perception in data annotation, as indicated by Cohen's Kappa scores. The developed hybrid detector-classifier model achieved an 85.50% mean average precision (mAP) at 0.5 intersections over union (IoU), tested with 402 images containing 561 sprout instances, with an inference time of 0.383 seconds per image. The results confirm the potential of the hybrid model as an alternative to current manual operations. This study serves as a practical case, for extensions to other applications within plant phenotyping communities.

Keywords: ginger seed germination, growth monitoring, deep learning, instance segmentation

1. Introduction

Ginger, Zingiber officinale Rosc., has been recognized as an important spice in the Asia and Africa region. It is a staple in everyday cuisine and is available in various forms, including powder, liquid extract, and oil. Besides its culinary uses, the abundance of gingerol and other bioactive compounds in ginger highlights its unique medicinal value [1]. The growing attention toward ginger has turned it into a commodity with high international demand. In 2019, countries such as India, Nigeria, China, and Nepal collectively produced more than 4.09 million metric tons of ginger, accounting for approximately 3.78 billion international dollars' worth of commodity [2]. Despite the rising demand, ginger production in some countries is not yet self-sustaining. For instance, ginger was reported to have one of the highest import dependency ratios among crops in Malaysia, reaching 81.5% in the year 2020 [3].

In ginger production, the life of a young plant is propagated through seed cuttings from the rhizome of mature ginger, usually referred to as "sett" or "bud". In traditional methods, after being cut from a rhizome, a ginger seed is buried directly in its growing medium. The survival of a ginger seed will only be examined at harvest after about 8 months of the planting period [4]. To better utilize resources, the common practice in ginger production is to cultivate seedlings from ginger seeds before planting, ensuring secure ginger germination [5]. Fig. 1 illustrates ginger propagation via a seed cut from a mature ginger rhizome. After being split from its mother rhizome, the buds on the seed gradually emerge and develop. The ginger seed is then harvested and planted in a nurturing medium to continue its growth and produce new mature rhizomes.

* Corresponding author. E-mail address: meyks@usm.my

English language proofreader: Yen-Chun Hsieh

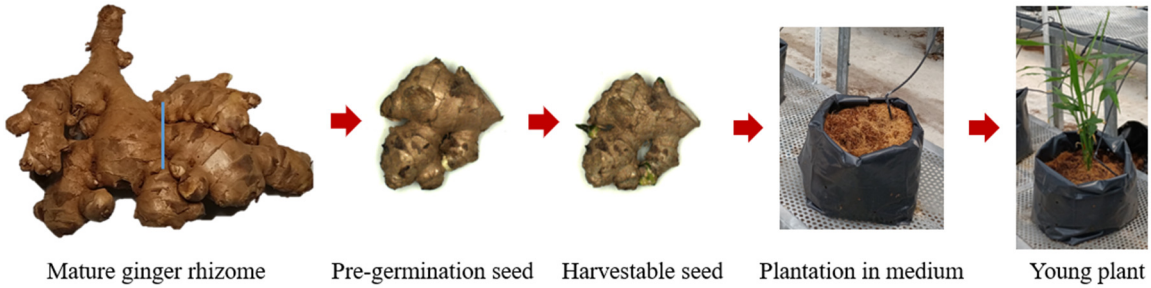


Fig. 1 Ginger propagation via seed from mature ginger rhizome

In any case, the germination of ginger seeds is crucial to ensure the efficient production of the underground plant. The condition of ginger seed sprouting has been shown to significantly contribute to the vigor of ginger plants and final yield, as demonstrated by Ai et al. [6]. Inspecting the growth status of ginger seed is therefore a cornerstone in ginger plantation. However, categorizing ginger seed growth status is challenging, as it requires not only long hours of work but also the expertise of a trained eye. The categorization is difficult to pin down with simple rules due to the irregular shapes and sizes. Meanwhile, the inspection process involving human experts lacks traceability. To date, there is no universal standard for ginger seed germination growth that can be used as a reference for inspection. On top of possible slip-throughs during judgment, the time-consuming manual inspection is sub-optimal for large-scale production. In the long run, the quantity and quality of ginger production will inevitably be constrained by the available manpower. Therefore, it is imperative to introduce an automated and intelligent solution, such as deep learning (DL), for rapid and traceable inspection of ginger seed growth.

Despite the use of DL methods in agriculture-related applications has been reported to be on the rise since 2016, the work to exploit the potential of DL methods in analyzing plant growth traits is still lacking in general compared to other applications in agriculture such as species classification, stress detection, and yield estimation [7]. Specifically, growth monitoring studies constitute only 14.08% of the 71 studies as reported by Yang and Xu [8]. Therefore, it is not surprising that the study related to ginger plant monitoring is nearly non-existence despite the recent interest in the crop. In a study to automate ginger shoot orientation recognition on a mobile platform, Fang et al. [9] demonstrated a successful application of the DL method for analyzing shoot orientation in the seed-sowing process. Nonetheless, there has not been a study on ginger germination growth monitoring to date. There is still a lack of investigation to obtain an effective object detector or feature extractor that can be applied to recognize growth stages of ginger seed. Therefore, this work aims to address the research gap by demonstrating the use of DL networks to detect, localize, and identify ginger sprouts of different growth stages in two-dimensional (2D) images.

The contributions of this work are as follows:

(1) This paper presents the first work to demonstrate the novel application of a DL network for ginger seed monitoring to three growth stages.

(2) The results in this paper reveal the potential of the DL network to make decisions that surpass human perception in growth stage classification at much higher speed. This highlights the applicability of the DL network for ginger seed monitoring applications.

(3) Lastly, this work contributes as a practical case study that utilizes a two-stage strategy in DL modeling, presenting a reference to other applications in plant phenotyping and computer vision communities.

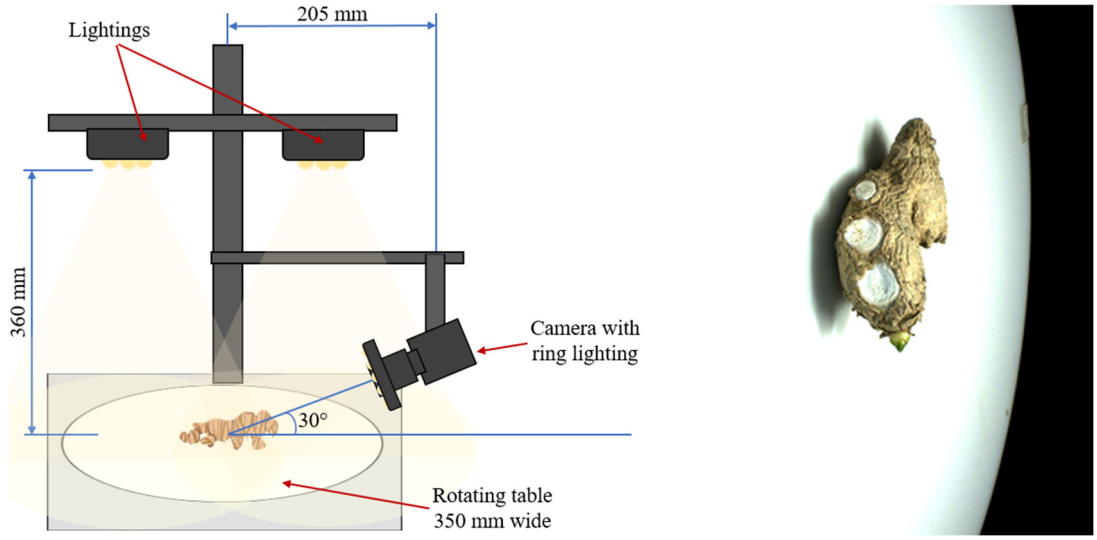
The following section of this paper provides an outline of details regarding materials, steps in acquisition, and the preparation of the dataset. Subsequently, the paper introduces a two-stage DL approach concept, followed by the presentation of the results from the model training work to select the best models for application. The performance of the selected DL model in the ginger seed germination task is then assessed and discussed in subsequent sections. Finally, the paper concludes with a summary and suggestions for future development.

2. Materials and Methods

This section outlines the data preparation in the study, which includes the equipment used, data categories, and an overview of the date set collected. Next, the strategy employed in this study is also explained, followed by the details on model training. Lastly, the categories of the detection result, as well as the performance metrics used are also detailed in this section.

2.1. Data preparation

In this work, a 1.3 megapixel (MP) color digital camera (HIKROBOT MV-CE013-50GC) was used with a 6 mm lens (OPTART MK-0614) to capture images of a ginger seed sample at 1280 pixels × 960 pixels resolution. The imaging system consisted of a 350 mm wide rotating table with a diffuse non-reflective white surface, a direct illumination source positioned 360 mm above the table, and a camera mounted at an oblique angle to the table, 30° from the imaging horizontal plane. Images of the specimen were captured at every 45° rotation to allow imaging of all sprouts that might emerge at any part of the seed. The imaging system used in this work is shown in Fig. 2.



(a) Imaging system (b) Example of the captured ginger seed image

Fig. 2 Imaging system and resultant image in this work

Raw mature ginger rhizomes of the species Zingiber officinale Rosc. were obtained from a local ginger plantation in Penang, Malaysia. The ginger specimens were prepared following industrial practice in the local ginger plantation. The gingers were cleaned with water to remove debris, etc. from their skin. After drying at room temperature, the ginger rhizomes were cut into seed pieces weighing from 20 g to 50 g. The ginger seeds were then allowed to germinate in a closed-area laboratory under a controlled environment with air at 26 °C and 70% relative humidity. To promote growth, the ginger setts were illuminated daily from 7 a.m. to 7 p.m. using white light-emitting diode (LED) lighting (TMS LITE HORTI HBL3-190-24V) with a color temperature of 5,300 K. In this study, 480 ginger seed samples underwent daily image acquisition for 21 days.

Throughout the acquisition period, only 282 ginger seed samples were found to have survived and demonstrated development. Besides, the collected images were filtered to remove those without visible seed sprouts. In the end, the dataset used in this study consists of 2,277 ginger seed sprout instances in 1,746 randomly selected images from the 282 ginger seed samples. The labeling of the collected dataset was done by a field expert with a background in ginger plantations. Regarding Ai et al. [6], ginger seed sprouts were categorized into three classes: Stage 1 (S1), Stage 2 (S2), and Stage 3 (S3), based on visual appearance, following the existing industrial practices. The regions of ginger seed sprouts in the images were manually identified and represented in polygon coordinates by the expert. The growth stages of ginger seed sprouts are illustrated in Fig. 3.

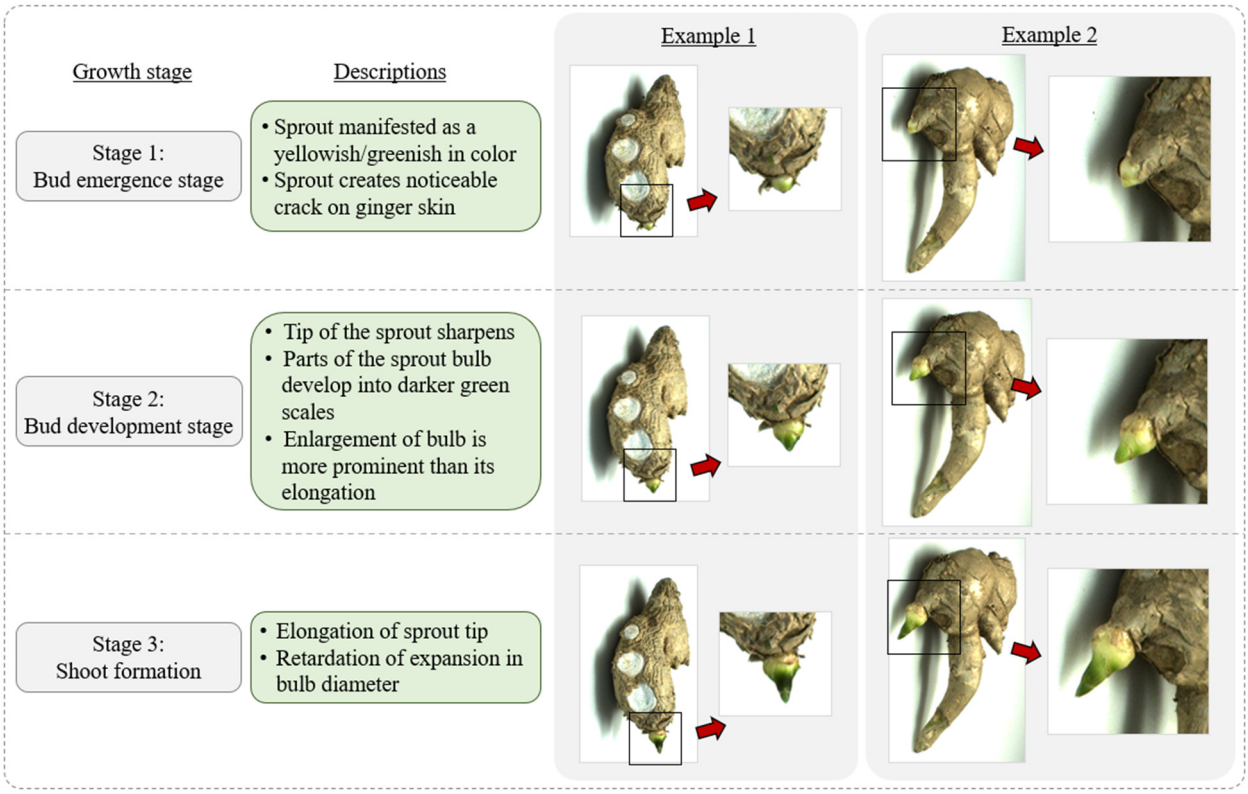


Fig. 3 Descriptions for the ginger seed growth stage and examples of images

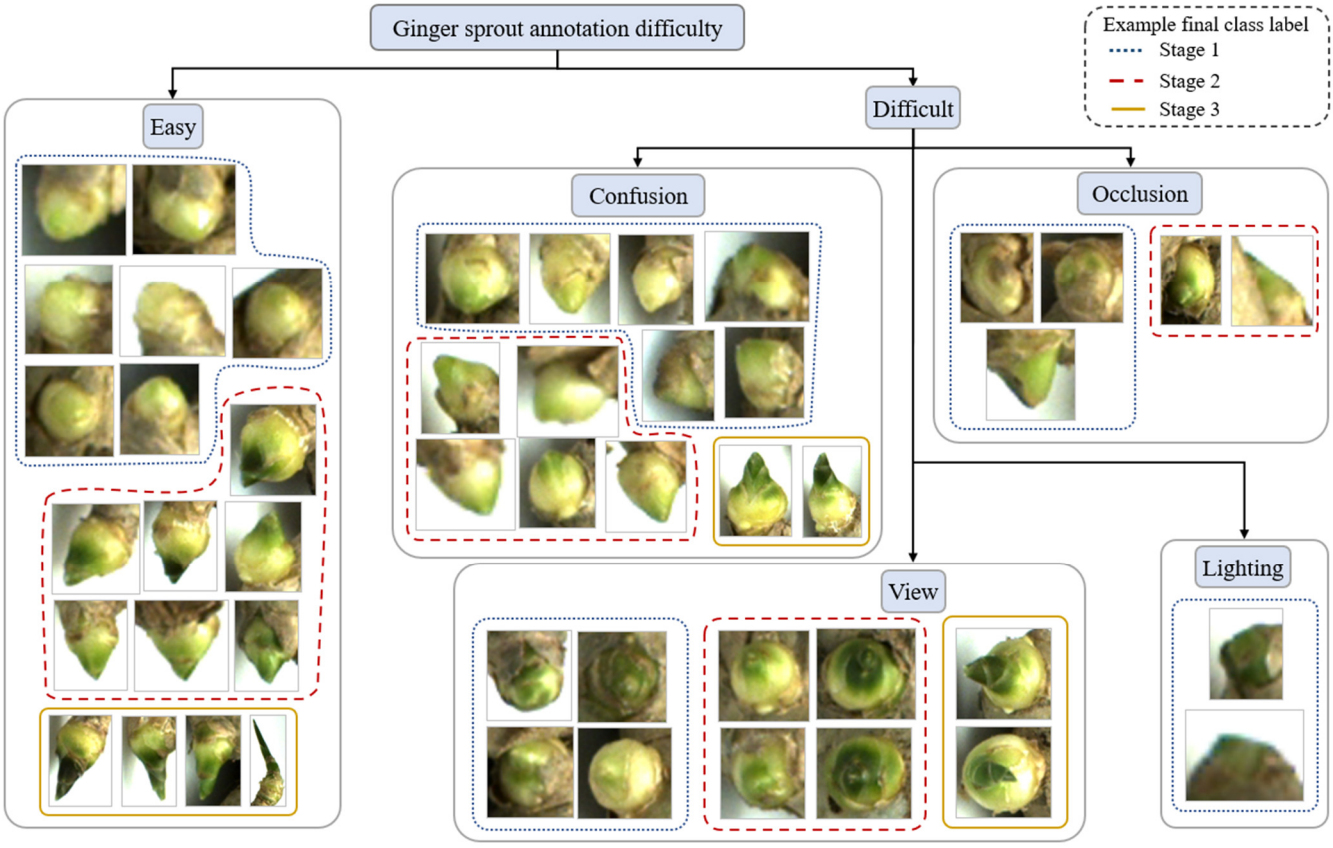


Fig. 4 Selected images of ginger sprouts with different difficulty annotations

For each identified region of sprout, additional labels were included besides its growth stage. These included a Boolean value indicating whether the working sample poses difficulty in labeling and its corresponding cause, attributed to either “confusion”, “occlusion”, “view”, or “lighting” by the expert. Due to the continuous nature of the plant growth, it may be challenging for the labeler to identify the subtle differences between two consecutive growth stages (“confusion”). Besides,

difficulty in labeling arises when the ginger seed sprout is blocked by other parts of itself ("occlusion") when it is imaged at a perpendicular angle ("view"), or under insufficient illumination ("lighting"). Several selected examples of "easy" and "difficult" ginger sprout instances are shown in Fig. 4. The images in Fig. 4 are scaled to the same height, maintaining their original aspect ratio.

Before further processing of the dataset, the labeling process in this work was repeated three times by the field expert. Cohen's Kappa statistic [10] was computed in this study to assess the reliability of the expert's annotations. The measure has a value ranging from -1 to +1, where -1 indicates complete disagreement, and +1 indicates perfect agreement. As shown in Table 1, a strong agreement (Kappa scores at 0.87) was observed for the growth stage class labels for the total 2,277 object instances. Despite the Kappa score of 0.57 indicating that the field expert wavered to decide whether a sample is easy or difficult, firmer decisions (Kappa scores at 0.79) were made on what caused the difficulties for the 750 (32.94%) samples marked "difficult" based on the majority of the annotations. It is undeniably challenging for the expert to classify seed growth stages.

Table 1 Cohen's Kappa scores for annotations on dataset characteristics

Annotation type	Sample size	Kappa score (Cohen 1960)
Class (3-class)	2277	0.8676
Difficulty (Easy/Difficult)	2277	0.5730
Difficulty causes (4 causes)	750	0.7873

The annotated dataset was then split into sets of 1,075 training images, 269 validation images, and 402 testing images (a ratio of 8:2:3). The split was carried out through stratification based on the possible seven ($2^3 - 1$) cases of the presence of object instances classes in the image. This was done to ensure that each training, validation, and testing set shares approximate similar object instance distribution across images for the multi-class detection task. Table 2 contains information about the dataset splits, including the number of images and the number of object instances.

Table 2 Dataset characteristics

Dataset split	No. of images	No. of instance	No. of instance by class			No. of instance by difficulty and cause					
						Easy	Difficult				
			S1	S2	S3		Confusion	View	Occlusion	Lighting	Total
Training	1075	1376	564	688	124	911	245	206	12	2	465
Validation	269	340	141	170	29	229	55	50	6	0	111
Testing	402	561	263	251	47	387	94	73	6	1	174
Total	1746	2277	968	1109	200	1527	394	329	24	3	750

In general, there was a considerably lower number of object instances with S3 class labels in the datasets. This can be attributed to a lower occurrence of a ginger seed developing into a later stage within the limited data collection period. Nonetheless, the random splitting, using the aforementioned stratification, resulted in approximately similar distributions of classes and difficulties for the three training, validation, and testing datasets.

Besides, it is evident from Table 2 that the difficulty in data annotation was mainly attributed to the causes of "confusion" and "view", which respectively accounted for 52.53% and 43.87% of "difficult" examples in the whole dataset used. This result reflects the substantial amount of confusion involved in the human expert's decisions during the classification task, besides the limitation caused by imaging settings.

2.2. Two-stage detector-classifier approach

In object detection or instance segmentation tasks, DL algorithms are commonly categorized into two types of architectures, namely, one-stage detectors such as single shot detectors (SSD) [11] and you only look once (YOLO) [12], and two-stage detectors, which are commonly represented by a region-based convolutional neural network (R-CNN) [13]. In general, R-CNN detectors are found to be more prominent than one-stage detectors in plant phenotyping-related research [14].

This choice can also be attributed to the fact that plant growth occurs eventually over time. Therefore, architectures based on the R-CNN family, with reported better detection accuracy, are preferred, since there is no pressing need for real-time detection [15]. Several studies have shown the advantages of a multi-stage network that benefits from the concept of task specialization. Separating a complex task into simpler tasks for different networks was reported to allow for more flexibility in models [16-17]. Compared to the knowledge necessary for fine-grained ginger seed sprout classification, the detection and localization of ginger seed sprouts in an image require relatively less expertise. Therefore, in this study, the multi-class ginger seed growth stage recognition task will be carried out in a two-stage approach. The concept of the two-stage approach is illustrated in Fig. 5.

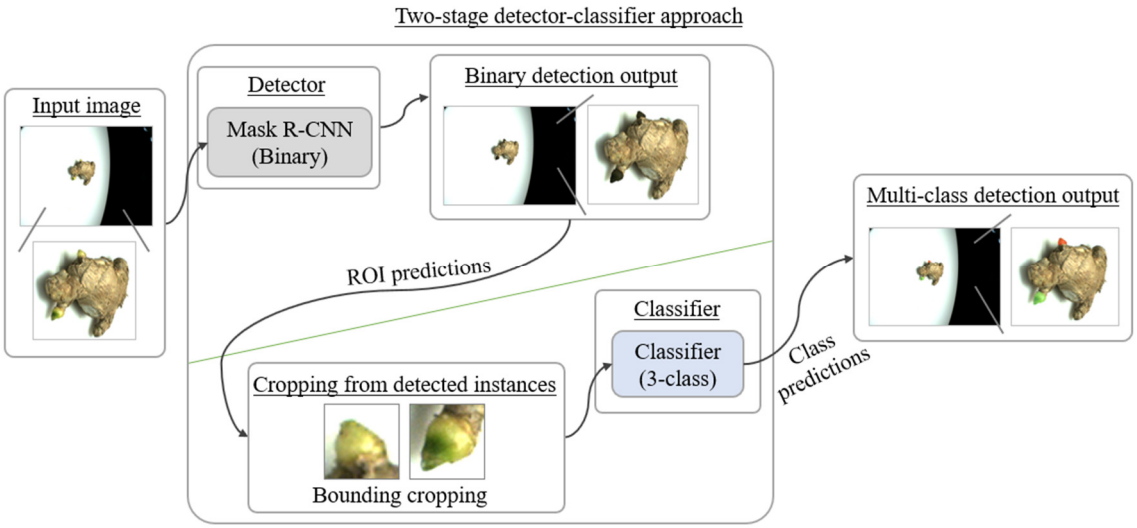


Fig. 5 Illustration of the proposed two-stage detector-classifier approach for multi-class instance segmentation

Firstly, Mask R-CNN [18] networks were developed and trained using transfer learning with different backbones for the detection of ginger sprouts regardless of their growth stages. Then, using the extracted image patches, a list of state-of-the-art networks with different capacities was trained and compared for classification. The optimal Mask R-CNN and classifier selected were then stacked for the ultimate multi-class instance segmentation task. As shown in Fig. 5, the detected sprout instances in images by the binary-class detector can be extracted based on their detected bounding boxes and fed to a multi-class classifier for further classification to growth stages. Lastly, the applicability of the hybrid model obtained was assessed using the testing dataset.

2.3. Model training

For binary-class Mask R-CNN training, several data augmentation techniques were applied to the dataset. The 1,075 images of the training set, consisting of 1,376 object instances, were further enlarged five times in number to 5,375 images and 26,328 object instances using a combination of pixel-level augmentation techniques and mosaic augmentation [19]. Firstly, each image in the training set was altered four times using at least one of the pixel-level techniques, including adjustments to image brightness, hue, saturation, gamma contrast, Gaussian noise, blur, rotation, and flipping. The parameters for these pixel-level augmentation techniques were selected to ensure that the alterations appeared reasonable, based on expert feedback. Then, mosaic augmentation was also carried out on both the original and pixel-level augmented images. Object instances from four randomly selected images were chosen each time to simulate the training images while maintaining the object scale.

For classifier training, image patches were extracted from the instance segmentation training and validation datasets based on the annotated bounding boxes, padded with a five-pixel border. This extraction step was applied directly to the validation dataset; however, the step was done on the training dataset after pixel-level augmentation but before mosaic augmentation. As a result, 6,880 and 340 image patches were extracted from the training and validation datasets, respectively, using the cropping method. Fig. 6 summarizes the aforementioned dataset preparation steps.

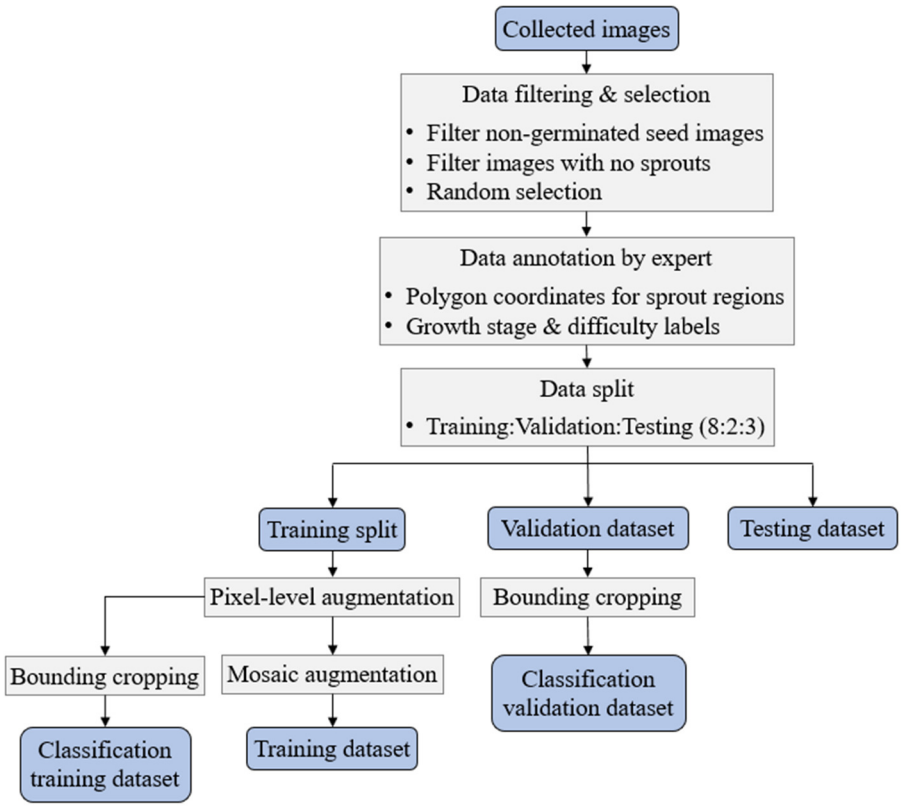
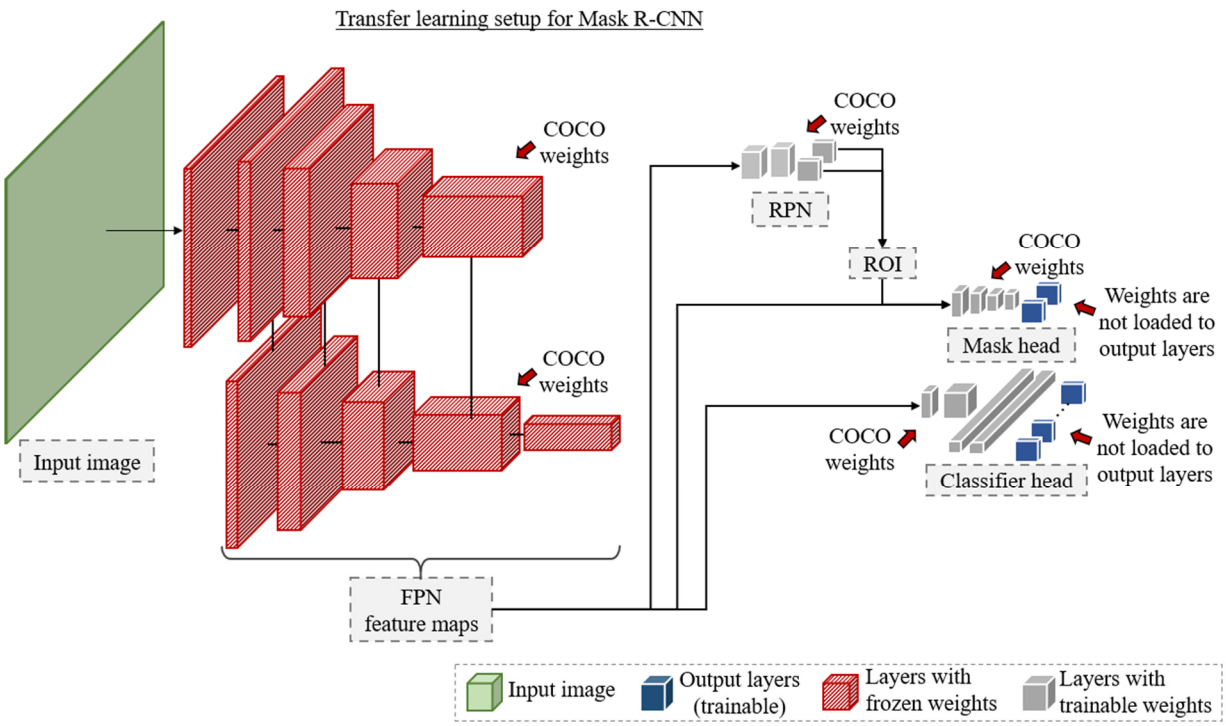


Fig. 6 Dataset preparation summary

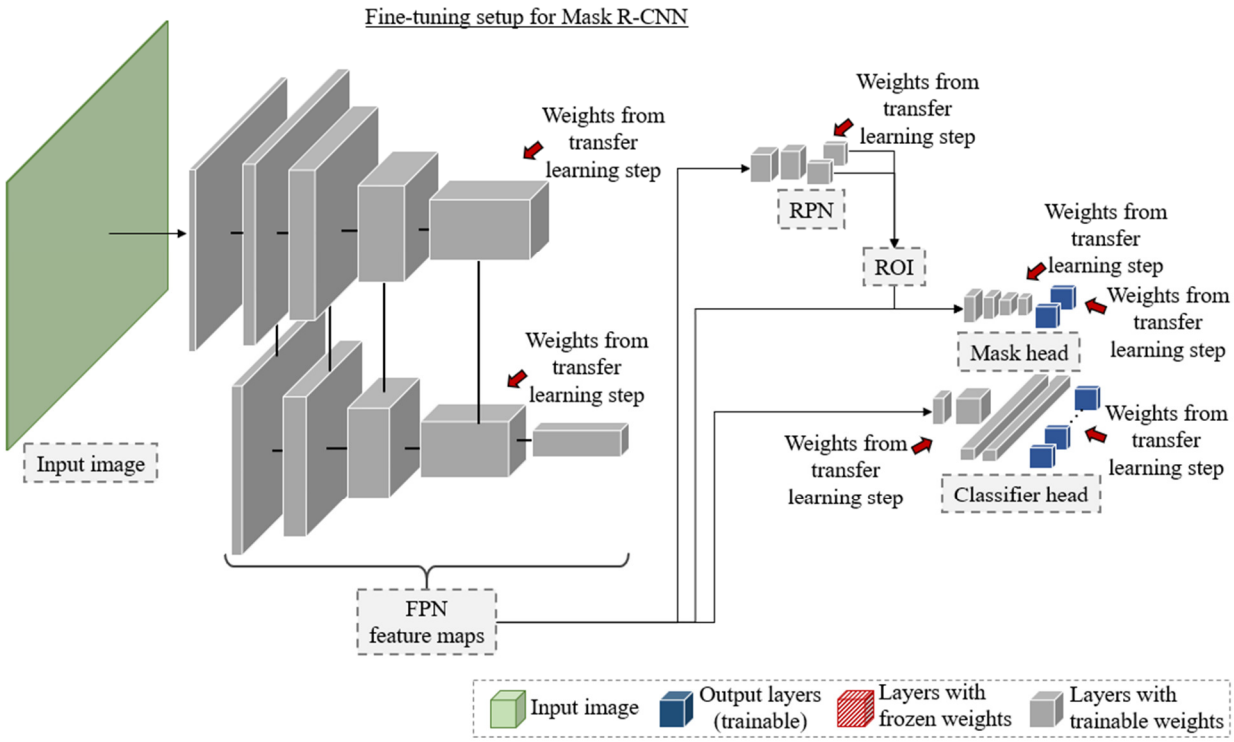
2.4. *Training details*

In this study, all models were trained using a machine equipped with an Intel i7-10700 CPU (16MB cache, 2.90 GHz) CPU, 32 GB DDR4 RAM modules, and an NVIDIA Quadro RTX4000 GPU with 8 GB GDDR6 RAM. In the first part of the experiment, the training of binary-class Mask R-CNN was carried out in two steps using the training dataset, as illustrated in Fig. 7.



(a) Illustration of transfer learning for Mask R-CNN

Fig. 7 Illustration of transfer learning and fine-tuning steps for Mask R-CNN training in this work



(b) Illustration of fine-tuning for Mask R-CNN

Fig. 7 Illustration of transfer learning and fine-tuning steps for Mask R-CNN training in this work (continued)

As shown in Fig. 7(a), during the transfer learning step, common objects in context (COCO) pre-trained weights were loaded for all layers in the network, except for the output layers. During transfer learning in this work, the weights for the layers up to the feature pyramid network (FPN) were held constant, while the weights in later parts of the network, including the region proposal network (RPN), mask head, and classifier head were made trainable.

Then, in the second fine-tuning step, the model obtained was subsequently fine-tuned with weight updates in all layers, as depicted in Fig. 7(b). In this step, the model was loaded with the weights obtained from the previous transfer learning step for all layers, including the output layers. The weights from all layers were set to be trainable in the fine-tuning step.

In both steps, the training dataset was used, and the early stopping of 10 epochs was applied to avoid overfitting the models. All the Mask R-CNN models were trained using a stochastic gradient descent (SGD) optimizer with a batch size of one image. The image was padded and resized to 1024 pixels × 1024 pixels for Mask R-CNN input. The Mask R-CNN was trained with two backbones (ResNet-50 and ResNet-101). The hyperparameters used in this study are depicted in Table 3.

Table 3 Binary-class Mask R-CNN hyperparameters for ginger seed sprout detection training

No.	Backbone	Training mode	Hyperparameters			Trained epochs
			Learning rate	Weight decay	Weights	
1	ResNet-101	Transfer learning	1×10^{-4}	1×10^{-4}	COCO	22
2		Fine-tuning	1×10^{-5}	1×10^{-3}	No. 1 last epoch	12
3	ResNet-50	Transfer learning	1×10^{-4}	1×10^{-4}	COCO	17
4		Fine-tuning	1×10^{-5}	1×10^{-3}	No. 3 last epoch	15

In the second part of the study, the classification training dataset was used to train image classifiers separately. A total of six state-of-the-art architectures with different capacities and complexities, indicated by their floating-point operations (FLOPs), were selected for training as multi-class classifiers in this study. These include DenseNet [20], EfficientNet [21], InceptionResNet [22], NASNet [23], ResNet, and Xception [24] which range from 4.29 G to 23.84 G FLOPs, as shown in Table 4.

Table 4 Hyperparameters for ginger seed sprout classification training

No.	Model name	FLOPs	Input image size (pixel)	Training mode	Batch size	Hyperparameter		Trained epochs
						L2 decay	Weights	
1	DenseNet201	4.29 G	224 × 224	Transfer learning	16	1×10^{-5}	ImageNet	198
				Fine-tuning		1×10^{-3}	No. 1 last epoch	102
2	EfficientNetB7	5.20 G	224 × 224	Transfer learning	16	1×10^{-5}	ImageNet	396
				Fine-tuning	8	1×10^{-3}	No. 2 last epoch	144
3	InceptionResNetV2	13.17 G	299 × 299	Transfer learning	16	1×10^{-5}	ImageNet	441
				Fine-tuning		1×10^{-3}	No. 3 last epoch	118
4	NASNetLarge	23.84 G	331 × 331	Transfer learning	16	1×10^{-5}	ImageNet	255
				Fine-tuning	4	1×10^{-3}	No. 4 last epoch	138
5	ResNet152V2	10.91 G	224 × 224	Transfer learning	16	1×10^{-5}	ImageNet	264
				Fine-tuning		1×10^{-3}	No. 5 last epoch	123
6	Xception	8.36 G	299 × 299	Transfer learning	16	1×10^{-5}	ImageNet	311
				Fine-tuning		1×10^{-3}	No. 7 last epoch	336

Similar to the previous part, the models were trained through transfer learning and fine-tuning steps, illustrated in Fig. 8. In the transfer learning step, ImageNet weights were loaded to the classifiers, excluding the output layers. All the layers before the output layers had frozen weights during transfer learning. Then, the weights obtained in the transfer learning step served as the initialization for fine-tuning, which involved weight updates in all layers. As shown in Table 4, Adam optimizer and dropout regularization with a rate of 0.5 was applied to all classifiers, with a 10-epoch early stopping setting. A learning rate of 1×10^{-5} was selected to reduce oscillation in optimization. The L2 regularization parameter increased from 1×10^{-5} during transfer learning to 1×10^{-3} for fine-tuning, introducing a higher level of regularization during fine-tuning with a large number of layers. The batch size was set at the maximum even number manageable by the GPU unit used in this study. To accommodate the training of the selected ImageNet-pretrained architectures listed in Table 4, the cropped patches were resized to the network's default input sizes.

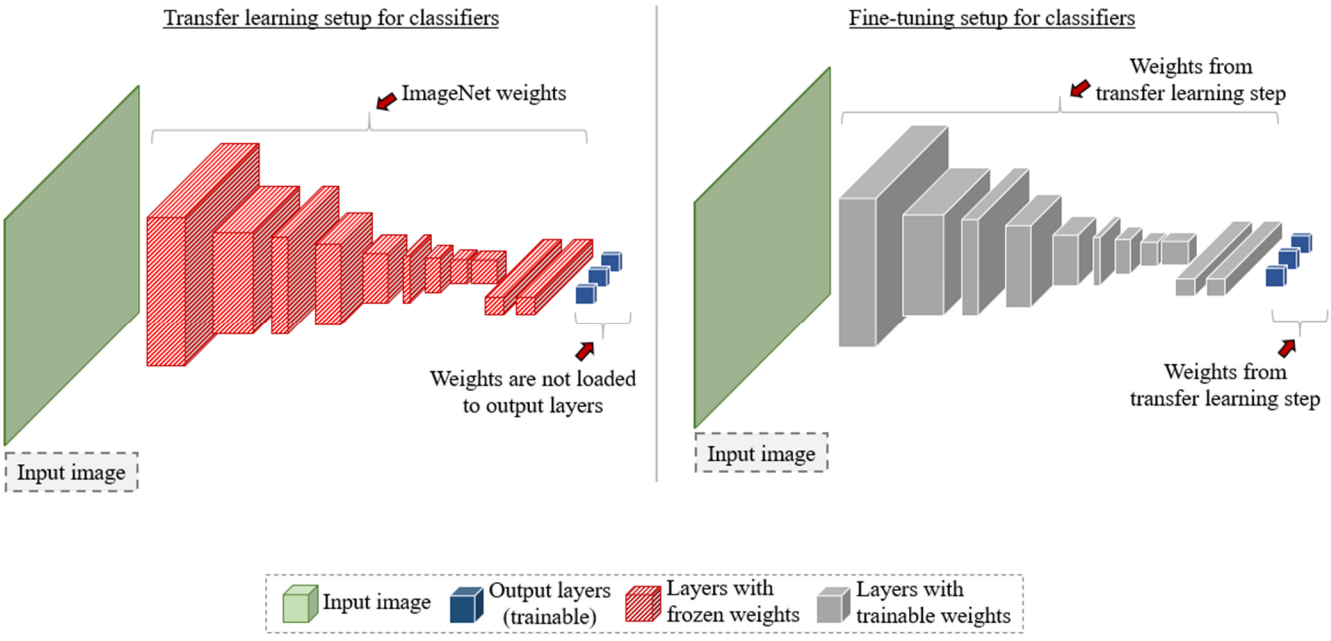


Fig. 8 Illustration of transfer learning and fine-tuning steps for classifiers training in this work

2.5. Performance evaluation

In instance segmentation, performance is evaluated based on a threshold of IoU to gauge the overlap between the predicted examples and the label examples. In this study, the IoU value was determined based on the mask predictions. A true positive (TP) prediction in instance segmentation is obtained when the prediction exhibits sufficient overlap with labeled

examples (i.e., greater or equal to an IoU threshold class) and has the same class label. In the event of duplicated predictions on the same labeled example, only the prediction with the highest confidence score is considered as the TP. Predictions that do not meet the TP criterion are considered false positives (FP).

An FP prediction can be further categorized into several types based on localization, classification performance, and duplication (DUP) in prediction. An FP prediction with at least 0.1 IoU for any labeled object in the image is regarded as a localization (LOC) error, while an FP prediction with an IoU value lower than 0.1 is considered an unlabeled background (BG) object [25]. Nevertheless, an FP prediction is also attributed to model confusion when an incorrect class prediction is made. Finally, when a labeled object instance is not detected by the model, the example is considered a false negative (FN).

$$Precision_i(x) = \frac{TP_i(x)}{TP_i(x) + FP_i(x)} \tag{1}$$

$$Recall_i(x) = \frac{TP_i(x)}{TP_i(x) + FN_i(x)} \tag{2}$$

$$Average\ precision,\ AP_{i,x} = \int_0^1 Precision_i(x) Recall_i(x) \tag{3}$$

$$Mean\ average\ precision,\ mAP_x = \frac{1}{n}\sum_i^n AP_{i,x} \tag{4}$$

The relationship between TP, FP, and FN, and average precision (AP) for a class i at the IoU threshold of x, is listed in Eq. (1) to (4). The primary indicator used for the detection performance is the mAP at 0.5 mask-based IoU, denoted as $mAP_{0.50}$. The computation of $mAP_{0.50}$ facilitates the evaluation of the detection abilities of the proposed model in sprout localization as well as the growth stages classification simultaneously. The integration of the area under the interpolated precision-recall curve for $mAP_{0.50}$ in this study was calculated using all points, as explained in Knausgård et al. [26].

$$F1\ score_i = \frac{2TP_i}{2TP_i + FP_i + FN_i} \tag{5}$$

$$Average\ F1\ score = \frac{1}{n}\sum_i^n F1\ score_i \tag{6}$$

In the three-class classification, the following rules were followed to assess the performance of the classifiers. For every class, a TP is accounted for when a prediction is made in the class correctly. If a prediction is made to a class incorrectly, it is considered an FP to the class. Similarly, when an example of a class is wrongly predicted as another class, it is considered a FN to the class. Then, the F1 score of a class was calculated using Eq. (5). The average F1 score across all classes, used as the indicator for classification performance, is given in Eq. (6).

3. Results and Discussion

In the early parts of this section, the results of the Mask R-CNN models and classification models are presented separately. Then, the multi-class detection results of the hybrid model, which is the combination of the best models selected in the aforementioned parts, are analyzed and discussed with illustration using selected examples.

3.1. Sprout detection using Mask R-CNN

The performances of the Mask R-CNN models using COCO-pre trained ResNet-101 and ResNet-50 backbones are depicted in Table 5 at their best $mAP_{0.50}$ epochs. From Table 5, the binary-class detection using Mask R-CNN recorded the best performance at the 11th epoch and 12th epoch with ResNet-101 and ResNet-50 backbone, respectively. The Mask R-CNN

architecture managed to achieve an outstanding performance of 96.17% $mAP_{0.50}$ in binary detection, identifying the presence of ginger seed sprouts in the image regardless of their growth stage. Therefore, Mask R-CNN with ResNet-50 backbone was selected for the binary-class detection task due to its superior $mAP_{0.50}$ measure on the validation set.

Table 5 Validation performances of the binary-class Mask R-CNN models

No.	Backbone	Selected epoch	Validation performance					
			Average IoU	$mAP_{0.50}$	$AP_{Easy,0.50}$	$AP_{Difficult,0.50}$	By difficulty cause	
							$AP_{Confusion,0.50}$	$AP_{View,0.50}$
1	ResNet-101	11	0.7050	0.9426	0.9432	0.9730	0.9455	1.0000
2	ResNet-50	12	0.7294	0.9617	0.9651	0.9820	0.9636	1.0000

3.2. Multi-class classification of growth stages

Table 6 depicts the performances of the classifiers trained in this study. In particular, the best classifier was the ResNet152V2 model, which attained an average 89.49% validation F1 score. Although ResNet152V2 has only a slight 0.01% advantage over InceptionResNetv2, which comes as second-best using the same cropping method, ResNet152V2 is favored due to its lower number of FLOPs.

Table 6 Validation performance of the classifiers trained in this study

No.	Model name	FLOPs	Selected epoch	Validation performance			
				F1 score (S1)	F1 score (S2)	F1 score (S3)	Avg. F1 score
1	DenseNet201	4.29 G	201	0.8935	0.8896	0.8889	0.8906
2	EfficientNetB7	5.20 G	417	0.9053	0.8974	0.8519	0.8848
3	InceptionResNetV2	13.17 G	481	0.8866	0.8855	0.9123	0.8948
4	NASNetLarge	23.84 G	262	0.8729	0.8709	0.8929	0.8789
5	ResNet152V2	10.91 G	271	0.8881	0.8841	0.9123	0.8949
6	Xception	8.36 G	362	0.8850	0.8802	0.8814	0.8822

Fig. 9 illustrates the losses as well as validation F1 scores computed during the two-step training of the best-performing Resnet152V2. In general, the losses of the trained models exhibited large spikes when transitioning from the transfer learning setting to the fine-tuning setting. This phenomenon can be attributed to the significantly lower number of layers being trained in the transfer learning stage, with only the last few layers included in training at this point. When the models were trained with the fine-tuning setting, a large increment in loss due to regularization was introduced. Nevertheless, this transition is also accompanied by improvement in classification, as reflected by the increments in the F1 score using the validation set. The learning that continued in the fine-tuning stage is reflected by the convergence of losses in the fine-tuning stage.

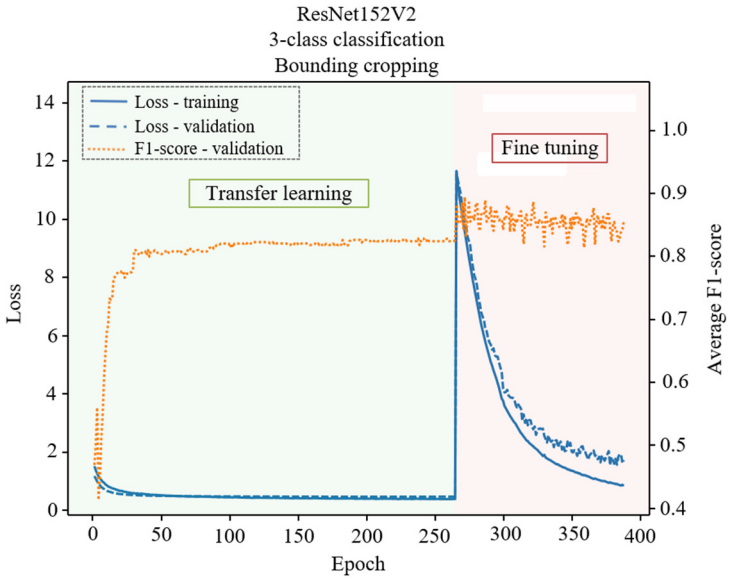


Fig. 9 Losses and validation F1 score for best-performing Resnet152V2 in the classification task

Due to its superior classification performance at lower computation cost, the ResNet152V2 bounding cropping model was selected to be employed as the classifier for the two-stage approach. An examination of the predictions on the validation set by the selected ResNet152V2 model using gradient-weighted class activation mapping (Grad-CAM) [27] revealed that the model has been sufficiently trained. Several selected examples of the Grad-CAM results are depicted in Fig. 10. The top row of images in Fig. 10 consists of the original images, while the middle row and bottom row consist of the corresponding Grad-CAM overlay images using ResNet152V2 model at the first epoch and the selected fine-tuned epoch, respectively. The "difficult" examples among the selection are depicted with blue borders in Fig. 10. Comparing the illustration for the model at the first training epoch and its selected fine-tuned state in Fig. 10, the gradient patterns shifted from dispersing patterns to appearing concentrated at certain salient parts of ginger sprouts. In general, predictions of S1 were associated with the whole region of the emerging sprout, while the regions of tips were revealed to be important in S2 and S3 predictions.

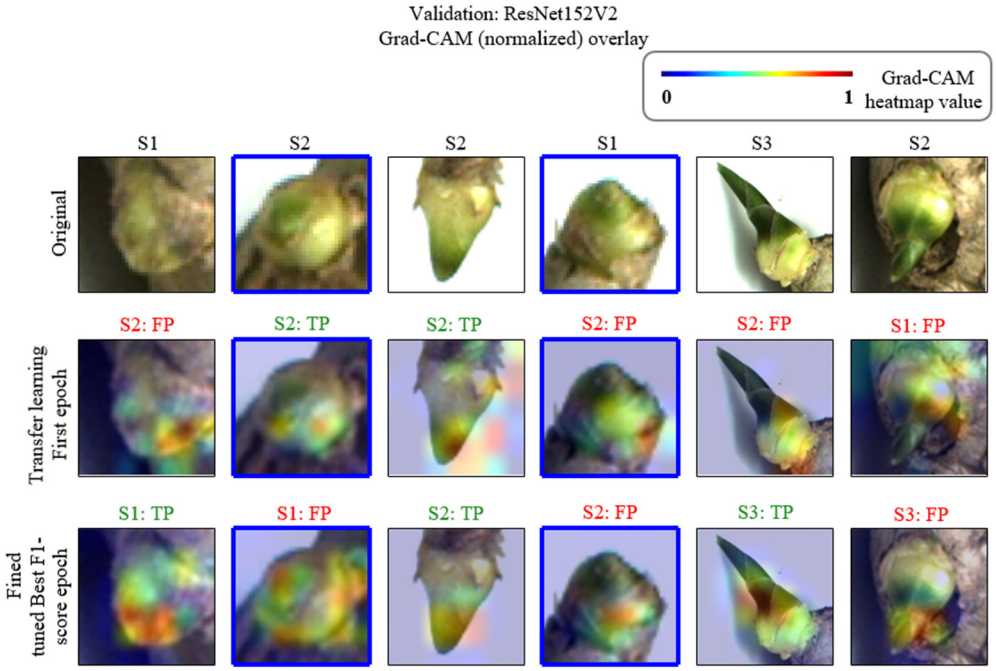


Fig. 10 Validation examples with Grad-CAM overlays (ResNet152V2)

3.3. Hybrid model (Binary-class detector + Multi-class classifier)

Table 7 depicts the testing performance of the binary-class Mask R-CNN and the hybrid model using the testing dataset, which consists of 402 images with 561 sprout instances. The binary-class Mask R-CNN managed to detect sprout instances regardless of growth stages at a $mAP_{0.50}$ of 93.10%. Coupling the Mask R-CNN with the selected bounding cropping ResNet152V2 classifier increases the model complexity, as reflected by the higher number of FLOPs at 320.19 G compared to the simpler binary-class Mask R-CNN at 309.28 G. However, the hybrid model achieved detection at a lower $mAP_{0.50}$ of 85.50%. Besides, the hybrid model marked the multi-class detection at an average inference time of 382.87 ms per image, which is an additional 94.08 ms compared to the binary-class Mask R-CNN. The increment in inference time, however, is insignificant.

Table 7 Average precision by class of the binary-class Mask R-CNN vs the hybrid model in testing

Detection type	Model	FLOPs	Avg. inference time/image (ms)	Avg. IoU	Average precision			
					By class			$mAP_{0.50}$
					$AP_{1,0.50}$ (S1)	$AP_{2,0.50}$ (S2)	$AP_{3,0.50}$ (S3)	
Binary-class	Mask R-CNN (ResNet-50)	309.28 G	288.79	0.7733	-	-	-	0.9310
Multi-class	Mask R-CNN (ResNet-50) + ResNet152V2	320.19 G	382.87	0.7733	0.7653	0.8444	0.9552	0.8550

Table 8 shows the $AP_{0.50}$ score by example difficulties for both models. From Table 8, the binary-class Mask R-CNN achieved 92.76% and 95.40% $AP_{0.50}$ for "easy" and "difficult" examples, respectively. Interestingly, when the similar binary-class Mask R-CNN was applied as a part of the hybrid model for multi-class detection, the measures dropped to 89.66% and 75.86% respectively. The comparable or even higher $AP_{0.50}$ score achieved for "difficult" examples in binary detection indicates that the Mask R-CNN architecture has managed to perform binary detection regardless of human intuition on the examples. In contrast, the larger number of mistakes made by the hybrid model on the "difficult" examples than the "easy" examples may also indicate a substantial level of agreement with human confusion in distinguishing sprout growth stages. In particular, the $AP_{0.50}$ observed for the "difficult" examples due to "confusion" was marked at only 71.28%. This result reflects the different nature of the task to detect the presence of the ginger seed sprout and the task to further classify them into different growth stages.

Table 8 Average precision by difficulty and attributed causes of the binary-class Mask R-CNN vs the hybrid model in testing

Detection type	Model	Average precision			
		By difficulty			
		$AP_{Easy,0.50}$	$AP_{Difficult,0.50}$	By difficulty cause	
				$AP_{Confusion,0.50}$	$AP_{View,0.50}$
Binary-class	Mask R-CNN (ResNet-50)	0.9276	0.9540	0.9681	0.9452
Multi-class	Mask R-CNN (ResNet-50) + ResNet152V2	0.8966	0.7586	0.7128	0.8356

Fig. 11 illustrates the confusion matrix of the multi-class detection normalized with the number of labeled examples in each class. It is noticeable that the confusion errors were concentrated between the two earlier S1 and S2 growth stages, while only a few errors were made for the S3 examples. Specifically, from Fig. 11, 96% of the S3 examples in the testing dataset were correctly detected compared to 83% and 86% for S1 and S2 respectively. Nonetheless, the results in the confusion matrix show that the classification mistakes by the model were mainly due to confusion between two consecutive growth stages. In contrast, there was no confusion between the S1 and S3. This finding is in line with the observations obtained by Wang et al. [28]. The authors have attributed the phenomenon to the higher difficulty in distinguishing the two growth stages that share higher similarities compared to earlier growth stages. In other words, this result also clearly reflects the sequential relationship in plant growth.

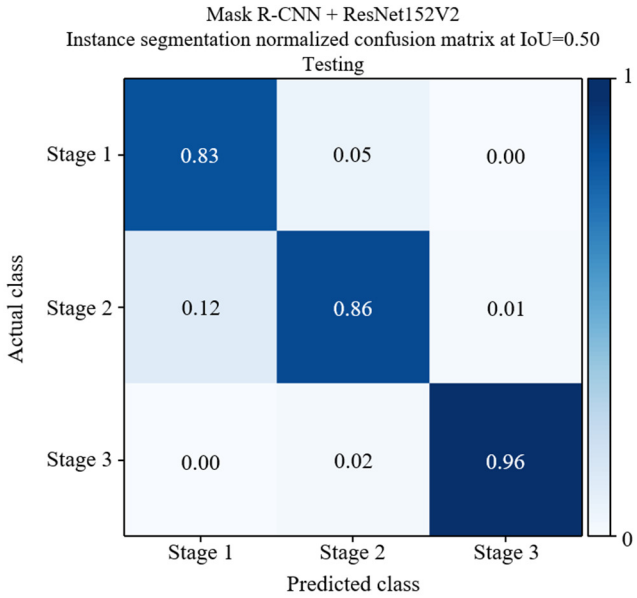


Fig. 11 Normalized confusion matrix: Mask R-CNN + ResNet152V2 (testing)

Besides, the predictions made by the two-stage hybrid model during testing were analyzed and depicted in Table 9. The results show that the two-stage model generally achieved more accurate detection in classes S2 and S3 than in class S1. According to Table 9, it is evident that the majority of the FP errors occurred in S1 examples. Out of the 267 positive S1

predictions made by the model, only 219 were predicted. This ratio is lower compared to S2 (215 TPs for 233 predicted positives) and S3 (45 TPs for 49 predicted positives) examples. In general, the FP errors made by the hybrid model were mainly due to confusion and localization errors. Specifically, confusion errors accounted for 60.42%, 77.78%, and 75% of the FP errors for each S1, S2, and S3 example, respectively.

Table 9 Multi-class detection results of the hybrid model in testing

Object class	No. of predicted positive	Detection results of object instances					
		TP	FP				FN
			Localization (LOC)	Duplication (DUP)	Confusion (CON)	Background (BG)	
S1	267	219	15	0	29	4	4
S2	233	215	4	0	14	0	18
S3	49	45	1	0	3	0	1

Another noteworthy observation from Table 9 is that the majority of the FN predictions made by the hybrid model are from the class S1, which was made up of 18 out of 19 of the FN predictions of the model. The investigation of the FN predictions revealed model weakness in the detection of small objects in the images. As shown in Fig. 12, it can be observed that FN detection tends to occur for object instances with smaller box areas. The misses in detection are concentrated for examples with less than 500 pixels of box area for the two-stage model. Nonetheless, this result indicates that FN predictions can be reduced using a higher-resolution camera.

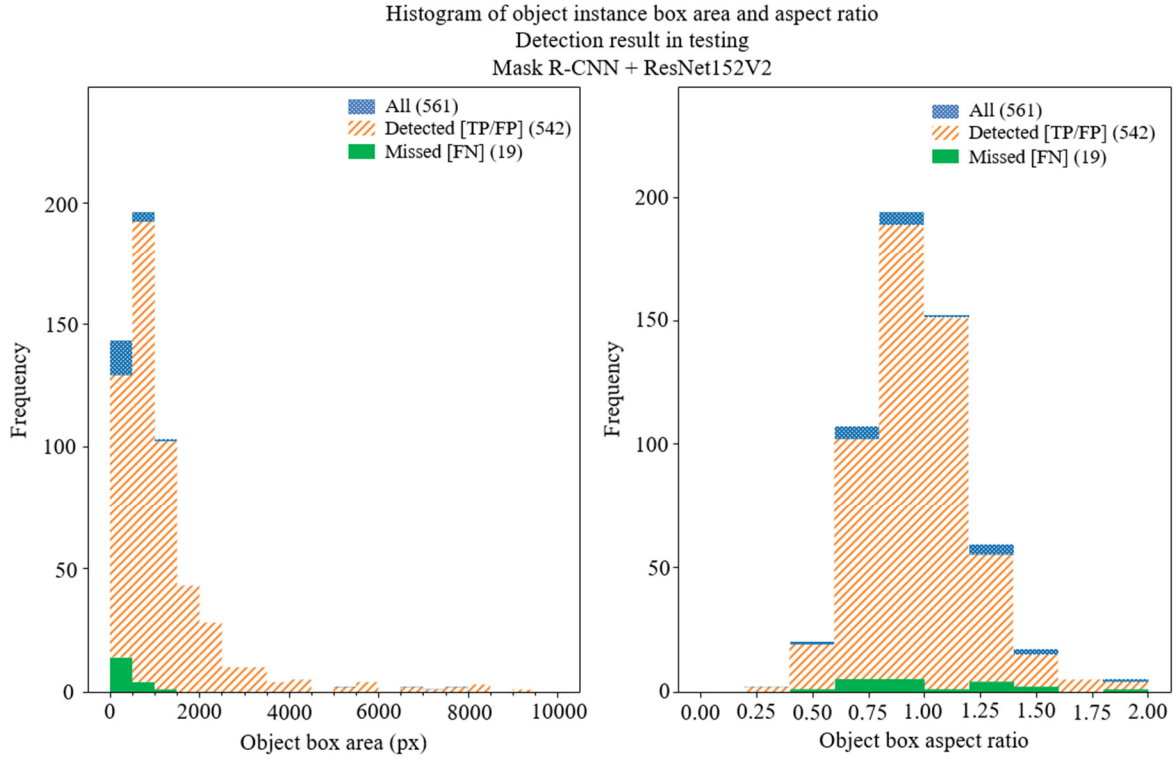
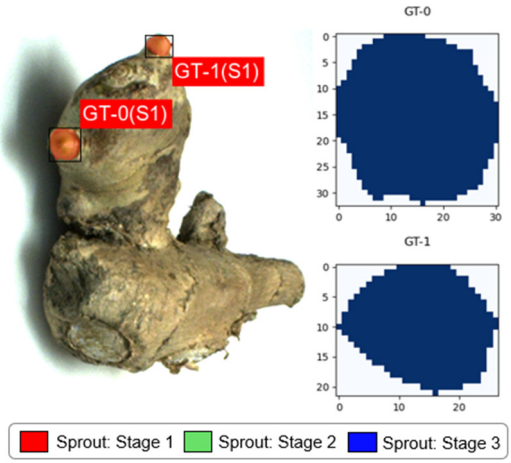
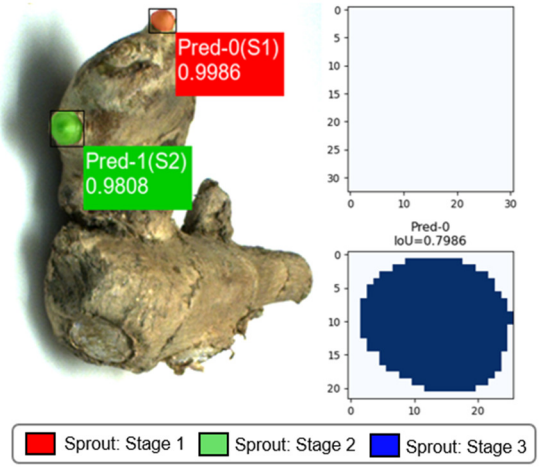


Fig. 12 Object instance box characteristics: Mask R-CNN + ResNet152V2 (testing)

Finally, Fig. 13, Fig. 14, and Fig. 15 each show a randomly selected example of the detections in the testing dataset. Fig. 13(a), Fig. 14(a), and Fig. 15(a) present the annotations by the experts for Examples A, B, and C, while Fig. 13(b), Fig. 14(b), and Fig. 15(b) depict the examples with prediction labels corresponding to Examples A, B and C using the hybrid model alongside confidence scores. Despite discrepancies between the segmented masks and the expert-labeled masks in these examples, the predicted masks by the hybrid model conform to the general shape of the sprouts, as reflected by the IoU scores, which average more than 77%. Regarding example A in Fig. 13, it is notable that the hybrid model suffers from confusion errors, where predictions are sufficiently detected and localized but not correctly categorized.

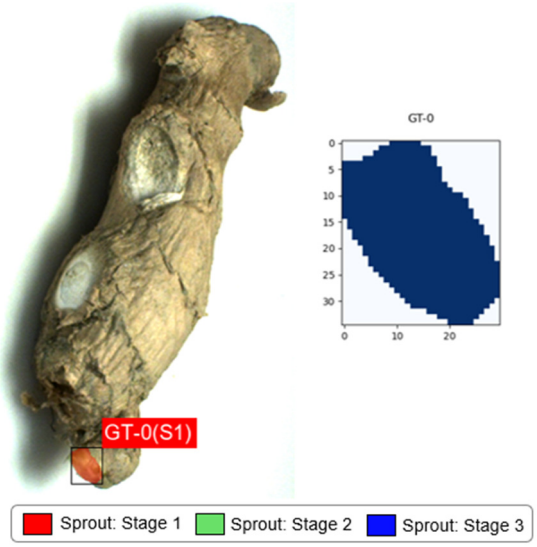


(a) Expert annotations in Example A

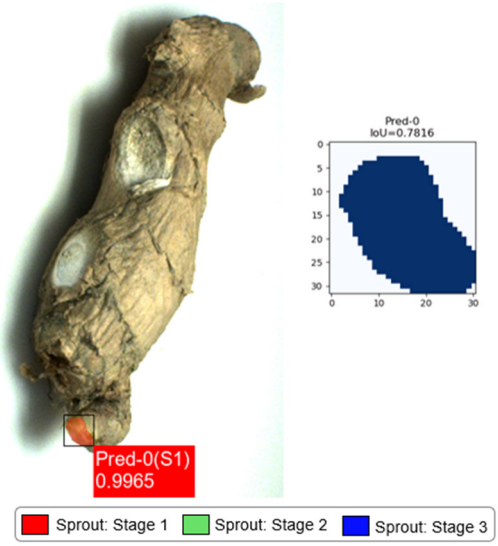


(b) Model predictions in Example A

Fig. 13 Annotated Example A in the testing dataset and corresponding predictions using the hybrid model

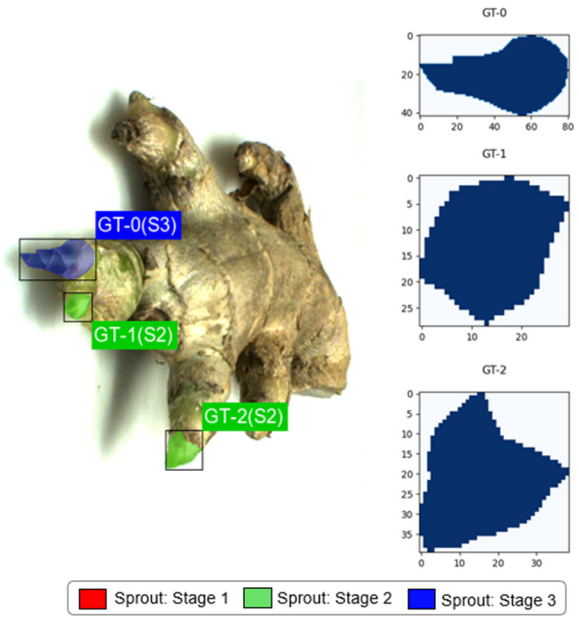


(a) Expert annotations in Example B

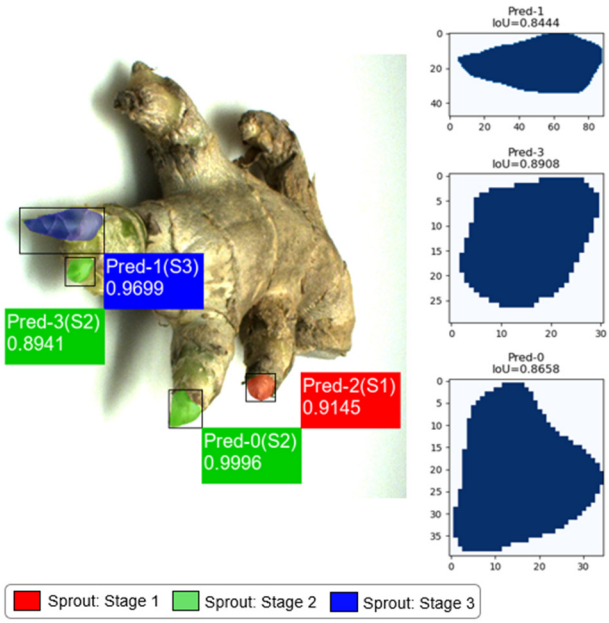


(b) Model predictions in Example B

Fig. 14 Annotated Example B in the testing dataset and corresponding predictions using the hybrid model



(a) Expert annotations in Example C



(b) Model predictions in Example C

Fig. 15 Annotated Example C in the testing dataset and corresponding predictions using the hybrid model

As indicated by the 0.87 Kappa score in Table 1, drawing a clear boundary between the growth stages of ginger seed sprouts has proven to be a daunting task for humans. In particular, the examples marked with "confusion" difficulty by the expert in the testing dataset were revealed to be the culprit in model performance. Nonetheless, the hybrid detector-classifier model managed to mark a detection performance of 85.50% $mAP_{0.50}$ in order of milliseconds per image. Furthermore, the examination of the model predictions revealed that the model decisions were in line with the human expert's confusion in growth stages classification. These results highlight the potential of the hybrid model as a rapid alternative to replace human inspection for ginger seed germination monitoring.

4. Conclusions

As the first work concerning the recognition of growth stages in ginger seed during germination, this study has successfully applied DL models to automate the recognition of three growth stages from images. This work also demonstrates the effectiveness of a two-stage strategy employing Mask R-CNN models for instance segmentation tasks. In future efforts, the incorporation of DL techniques, such as label smoothing, may be considered to mitigate the impact of human errors in annotation on model performance. At the same time, the results indicate that the exploration to benefit from the existing sequential relationship between object classes could be a worthwhile subject in the future of ginger seed monitoring. Another prospective research direction involves utilizing temporal information in a series of images. The a priori sequential relationship between plant growth stages, as demonstrated by Samiei et al. [29], may be exploited for classification.

Fundings

This work was supported by TMS LITE Sdn. Bhd. and Universiti Sains Malaysia. It was also partially funded by the Ministry of Higher Education Malaysia for the Fundamental Research Grant Scheme with Project Reference Code: FRGS/1/2021/ICT02/USM/02/2.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] K. Srinivasan, "Ginger Rhizomes (Zingiber Officinale): A Spice with Multiple Health Beneficial Potentials," PharmaNutrition, vol. 5, no. 1, pp. 18-28, March 2017.

[2] Food and Agriculture Organization of the United Nations, "FAOSTAT," https://www.fao.org/faostat/en/#data, April 22, 2022.

[3] Department of Statistics Malaysia, "Supply and Utilization Accounts Selected Agricultural Commodities 2016-2020," https://www.dosm.gov.my/v1/index.php?r=column/cthemeByCat&cat=164&bul_id=cHgwanhNdU4vWXRvc3pnZU9xSjZTUT09&menu_id=Z0VTZGU1UHBUT1VJMFlpaXRRR0xpdz09, August 26, 2021.

[4] K. P. P. Nair, The Agronomy and Economy of Turmeric and Ginger, 1st ed., Oxford: Elsevier, 2013.

[5] D. Saravanakumar, A Guide to Good Agricultural Practices for Commercial Production of Ginger under Field Conditions in Jamaica, Kingston: Food and Agriculture Organization of the United Nations, 2021.

[6] X. Ai, J. Song, and X. Xu, Ginger Production in Southeast Asia, 1st ed., Boca Raton: CRC Press, 2004.

[7] Y. S. Tong, T. H. Lee, and K. S. Yen, "Deep Learning for Image-Based Plant Growth Monitoring: A Review," International Journal of Engineering and Technology Innovation, vol. 12, no. 3, pp. 225-246, June 2022.

[8] B. Yang and Y. Xu, "Applications of Deep-Learning Approaches in Horticultural Research: A Review," Horticulture Research, vol. 8, article no. 123, 2021.

[9] L. Fang, Y. Wu, Y. Li, H. Guo, H. Zhang, X. Wang, et al., "Ginger Seeding Detection and Shoot Orientation Discrimination Using an Improved YOLOv4-LITE Network," Agronomy, vol. 11, no. 11, article no. 2328, November 2021.

[10] M. L. McHugh, "Interrater Reliability: The Kappa Statistic," Biochemia Medica, vol. 22, no. 3, pp. 276-282, 2012.
[11] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, et al., "SSD: Single Shot MultiBox Detector," Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, pp. 21-37, October 2016.
[12] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," IEEE Conference on Computer Vision and Pattern Recognition, pp. 779-788, June 2016.
[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," IEEE Conference on Computer Vision and Pattern Recognition, pp. 580-587, June 2014.
[14] A. Abade, P. A. Ferreira, and F. de Barros Vidal, "Plant Diseases Recognition on Images Using Convolutional Neural Networks: A Systematic Review," Computers and Electronics in Agriculture, vol. 185, article no. 106125, June 2021.
[15] N. Genze, R. Bharti, M. Grieb, S. J. Schultheiss, and D. G. Grimm, "Accurate Machine Learning-Based Germination Detection, Prediction and Quality Assessment of Three Grain Crops," Plant Methods, vol. 16, article no. 157, 2020.
[16] K. M. Knausgård, A. Wiklund, T. K. Sørdalen, K. T. Halvorsen, A. R. Kleiven, L. Jiao, et al., "Temperate Fish Detection and Classification: A Deep Learning Based Approach," Applied Intelligence, vol. 52, no. 6, pp. 6988-7001, April 2022.
[17] C. Sandoval, E. Pirogova, and M. Lech, "Two-Stage Deep Learning Approach to the Classification of Fine-Art Paintings," IEEE Access, vol. 7, pp. 41770-41781, March 2019.
[18] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 2, pp. 386-397, June 2018.
[19] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," https://doi.org/10.48550/arXiv.2004.10934, April 23, 2020.
[20] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," IEEE Conference on Computer Vision and Pattern Recognition, pp. 2261-2269, July 2017.
[21] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," International Conference on Machine Learning, PMLR, vol. 97, pp. 6105-6114, June 2019.
[22] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, Inception-Resnet and the Impact of Residual Connections on Learning," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31, no. 1, article no. 11231, 2017.
[23] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning Transferable Architectures for Scalable Image Recognition," IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8697-8710, June 2018.
[24] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," IEEE Conference on Computer Vision and Pattern Recognition, pp. 1800-1807, July 2017.
[25] D. Hoiem, Y. Chodpathumwan, and Q. Dai, "Diagnosing Error in Object Detectors," Computer Vision – ECCV 2012: 12th European Conference on Computer Vision, pp. 340-353, October 2012.
[26] R. Padilla, S. L. Netto, and E. A. B. da. Silva, "A Survey on Performance Metrics for Object-Detection Algorithms," International Conference on Systems, Signals and Image Processing, pp. 237-242, July 2020.
[27] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," IEEE International Conference on Computer Vision, pp. 618-626, October 2017.
[28] S. Wang, Y. Li, J. Yuan, L. Song, X. Liu, and X. Liu, "Recognition of Cotton Growth Period for Precise Spraying Based on Convolution Neural Network," Information Processing in Agriculture, vol. 8, no. 2, pp. 219-231, June 2021.
[29] S. Samiei, P. Rasti, J. Ly Vu, J. Buitink, and D. Rousseau, "Deep Learning-Based Detection of Seedling Development," Plant Methods, vol. 16, article no. 103, 2020.