
Article:
Su, J., Anderson, S., Javed, M. et al. (3 more authors) (2023) Plant leaf deep semantic segmentation and a novel benchmark dataset for morning glory plant harvesting. Neurocomputing, 555. 126609. ISSN 0925-2312

https://doi.org/10.1016/j.neucom.2023.126609

Neurocomputing 555 (2023) 126609

Contents lists available at ScienceDirect

Neurocomputing

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

Plant leaf deep semantic segmentation and a novel benchmark dataset for morning glory plant harvesting

Jingxuan Su [a,*], Sean Anderson [a], Mahed Javed [a], Charoenchai Khompatraporn [b], Apinanthana Udomsakdigool [b], Lyudmila Mihaylova [a]

[a] Department of Automatic Control and Systems Engineering, The University of Sheffield, Sheffield S10 3JD, United Kingdom
[b] Department of Production Engineering, King Mongkut's University of Technology Thonburi (KMUTT), Bangkok, Thailand

ARTICLE INFO

Keywords:
Semantic image segmentation
Yield prediction
Precision agriculture
Deep learning methods
Encoder–decoder networks
Dense block

ABSTRACT

Computer vision and deep learning have made substantial progress in the areas of agriculture and smart farming, particularly for enhancing crop production using image segmentation techniques for crop yield prediction. Further improvements to crop yield prediction results can be achieved by developing accurate and efficient methods. In response to such demands, this paper proposes a novel convolutional neural network architecture, called densely connected SegNet (D-SegNet) and demonstrates its advantages on plant segmentation using a new morning glory plant dataset, and also on a complimentary publicly available dataset to promote research in this direction. The D-SegNet is evaluated using 10-fold cross validation. It achieves performance better than the state-of-the-art SegNet algorithm. The evaluated precision, recall and F1-score values are 98.20%, 90.64% and 94.26%, respectively, for the morning glory plant dataset. The intersection over union (IoU) value in the image segmentation tasks is 90.56%. A series of experiments on the morning glory plant dataset as well as on the publicly available dataset were conducted. The results show that the proposed method achieves accurate segmentation results and can be useful for assessing the plant weight during harvesting. In summary, this new plant segmentation network, D-SegNet, could form an important component of future cloud-based machine learning systems to predict crop yield from noisy smartphone images taken in the field.

1. Introduction

As the global population increases it is vital to tackle the challenges of agricultural production to improve food security and make production efficient [1]. Smart farming [2] is one promising area that supports the growth of agriculture production and its efficiency by leveraging technological advances in sensing, monitoring, and artificial intelligence [3,4]. Within the broad area of smart farming, crop planning, monitoring, and yield prediction [5] have been considered some of the dominant challenges in promoting coordinated efficiency and maximizing the economic potential of modernizing agriculture.

Yield prediction is one particularly important aspect of an agricultural business for the following reasons. The buyers of fresh products often plan the procurement based on the sale estimates or customers' orders. Suppose the buyers are aware that there will be a shortage or excess of production from the farmers. In that case, they can plan ahead of time by seeking additional products from retailers and buying or selling the excess to other buyers. By gaining more accurate yield information, the buyers can manage their incoming stock effectively. This can result in keeping the price for the farmers high and at the same time the management cost for the buyers low. With an accurate yield prediction, the buyer of the product can save costs through better planning their sourcing strategy and simultaneously better serve the needs of the customers. This further promotes the buyer's business, leading to close collaboration between farmers and buyers.

This paper presents a solution for crop yield prediction using computer vision and image processing methods. Image processing methods can assist precision agriculture, for example in plant phonology [6], automatic segmentation of leaf images [7,8] and yield prediction [4]. However, current methods for segmentation are not sufficiently accurate to be reliable and useful in yield prediction [9], especially using simple methods such as k-means clustering [10]. There is a wealth of machine learning methods for image analysis [11,12] — from support vector machines (SVM) [13] and artificial neural networks (ANN) [14] to deep learning (DL) [15]. The conventional neural

* Corresponding author.
E-mail addresses: jsu14@sheffield.ac.uk (J. Su), s.anderson@sheffield.ac.uk (S. Anderson), mahed95@yahoo.com (M. Javed), charoenchai.kho@kmutt.ac.th (C. Khompatraporn), apinanthana.udo@kmutt.ac.th (A. Udomsakdigool), l.s.mihaylova@sheffield.ac.uk (L. Mihaylova).

https://doi.org/10.1016/j.neucom.2023.126609
Received 5 August 2022; Received in revised form 13 July 2023; Accepted 24 July 2023
Available online 28 July 2023

learning methods [16] mentioned above aim at solving different tasks, including object detection and segmentation. Due to the accuracy demands of yield prediction, image-based, semantic segmentation at a pixel level is an appealing approach and is becoming popular in this application domain [17,18]. However, standard convolutional neural networks [19] face often problems such as slow convergence, and loss of vital feature information when using several convolutional layers, leading to inaccurate segmentation results. Thus, an optimized network architecture needs to be proposed. At present, most semantic segmentation methods conduct research on expanding the depth [20,21] or width [22,23] of network architectures. Deep learning methods [24, 25], as data-driven methods, provide efficient solutions for big data and automated feature extraction, with high performance and accuracy [12, 26].

In this paper, we propose a new type of semantic segmentation architecture for yield prediction, which comprises encoder–decoder [27] and dense block [28] structures. The encoder is comprised of thirteen convolutional layers of a VGG-16 network [20]. Each encoder is linked to a decoder and hence there are thirteen corresponding decoders. Typically, the convolutional layers of the encoder comprise batch normalization and rectified linear unit (ReLU) non-linear operations, followed by non-overlapping max pooling and sub-sampling layers. The sparse encoding due to the pooling process is up-sampling in the decoder using the max pooling indices in the encoding sequence. This has the important advantage of retaining class boundary details in the segmented images and also reducing the total number of model parameters. The model is trained end to end using the stochastic gradient descent method [29].

The structure of a dense block includes a concatenation between layers. The concatenation enhances feature reuse, which directly connects every layer. This connection pattern not only propagates features to the next layer but to every layer in a dense block. We propose a novel deep learning architecture here for dense semantic segmentation, which extends the standard SegNet architecture with dense blocks, hence we call this architecture D-SegNet. We demonstrate in this paper that this new type of architecture outperforms the state-of-the-art algorithms and achieves high accuracy on crop yield prediction.

The main contributions of this paper are the following:

1. A new deep learning architecture, called D-SegNet is proposed, where the new feature is the use of dense blocks to augment the standard SegNet architecture. A concatenation between the layers in the dense block is introduced. The main advantage of D-SegNet consists of the improved feature maps extracted from the images.
2. A new and comprehensive dataset for the morning glory plant is collected by us and made public, which is beneficial for reproducible research. The data [30,31], available on GitHub, contains original and mask images, which are ready for training and testing of algorithms for semantic image segmentation.
3. Experimental analysis on the morning glory plant dataset and ImageCLEF (Pl@ntleaves) dataset [32] using the proposed approach and other semantic segmentation techniques.
4. The performance of D-SegNet algorithm is evaluated and thoroughly validated over several metrics such as precision, recall, F1-score and Intersection over Union (IoU). Comparative results with the standard SegNet architecture are presented.

The rest of this paper is organized as follows. The D-SegNet is described in detail with the network architecture graph, parameters and layer connections in Section 2. The morning glory plant dataset and semi-automatic labelling method are described in Section 3. Section 4 presents the experimental analysis with different metrics. Finally, Section 5 summarizes the results and assessment.

2. The D-SegNet architecture

This section presents a new architecture called D-SegNet for semantic segmentation, which augments the encoder–decoder architecture of SegNet [27] with dense blocks [28]. The D-SegNet architecture combines the improved capability of feature extraction from the dense block with the computational efficiency of the SegNet encoder–decoder architecture, including efficient memory use and fast computation.

2.1. Encoder–decoder architecture

The encoder–decoder architecture of D-SegNet is illustrated in Fig. 1. The encoder consists of five blocks which include 'Convolution (Conv)+Batch Normalisation (BN)+ Rectified Linear (ReLU)' operations and they form the Downsample Blocks. In each Downsample Block, a convolutional layer uses a 3×3 kernel and a stride equal to 1. Batch normalization is then applied to the output of the convolutional layer. Meanwhile, the max pooling with 2×2 window and stride 2 achieve translation invariance. Therefore, the size of feature maps is changed regularly from 224×224 to 7×7 after the image is passed through five Downsample Blocks. The role of the encoder is to generate feature maps with semantic information.

The decoder which contains five Upsample Blocks upsamples the feature maps from the encoder output by using the memorized max-pooling indices to produce sparse feature maps. The size of these sparse feature maps is rescaled to the size of the original image. The purpose of the rescale operation in the decoder is to map the sparse feature maps to the input image to implement pixel-by-pixel classification.

2.2. Dense block

Recent CNNs benefit from very deep convolution layers to capture rich feature representations, due to the fact that 'the deeper the network, the better [33]. However, training deep neural networks may need a huge amount of time and computational resources, due to redundant feature maps. The dense block proposed in [28] is considered an effective method to address this challenge. It encourages feature reuse and makes it efficient to train a very deep network. The dense block improves the flow of features throughout the network by connecting all layers with each other. In addition, it also enhances feature propagation. The dense connectivity in a dense block can be formulated as follows:

$$x_d = H_d\left(\left[x_0, x_1, \ldots, x_{d-1}\right]\right), \tag{1}$$

where x_0 to x_d are feature maps from layer 0 to layer d in a dense block, respectively. Here $\left[x_0, x_1, \ldots, x_{d-1}\right]$ refers to the concatenation of feature maps from x_0 to x_{d-1}, H_d denotes a non-linear transformation, including BN [34], ReLU, and Conv operations. The BN size is set up to 4 to keep a lightweight network. In this way, each layer within a dense block has direct connections with all subsequent layers, as shown in Fig. 2. According to (1), the channel number of feature maps in the dth layer of each dense block is $k_0 + k \times (d-1)$, where k_0 is the number of input channels. The growth rate is termed k, following the same notation as in [28]. In [28], due to the feature map size, k is set up equal to 32. Each dense block has 4 layers with a growth rate of $k = 32$. Table 1 gives all parameters of the D-SegNet algorithm.

2.3. Optimal semantic segmentation model (D-SegNet)

The D-SegNet as a novel architecture has an outstanding framework and computational capability. This powerful segmentation engine consists of a deep convolutional encoder–decoder architecture, dense blocks, and a pixel-level classification layer. Table 1 lists the details of D-SegNet architecture. From Table 1, Dense Block is directly connected to Downsample/Upsample Blocks. Specifically, 1×1 $conv$ layers are set before the 3×3 $conv$ layers to avoid increasing the model parameters.

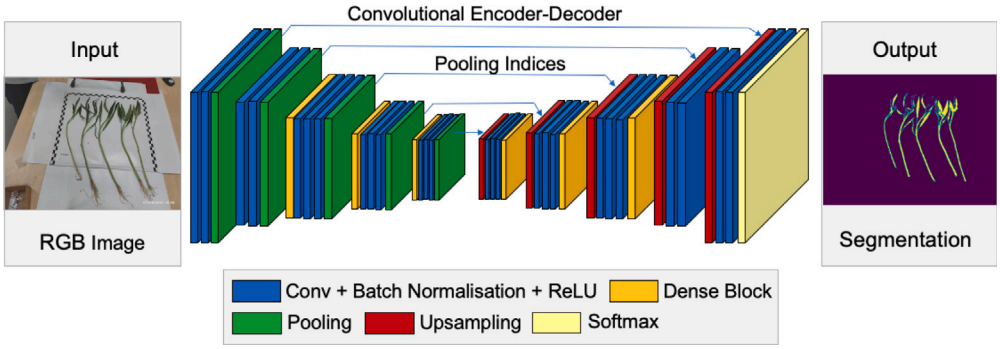

Fig. 1. A schematic of the D-SegNet architecture. The encoders are based on the 13 convolutional layers of the VGG-16 network [20]. The decoder places corresponding layers in reverse. The key novel feature compared to SegNet are the Dense blocks, which are added after each encoder or decoder section.

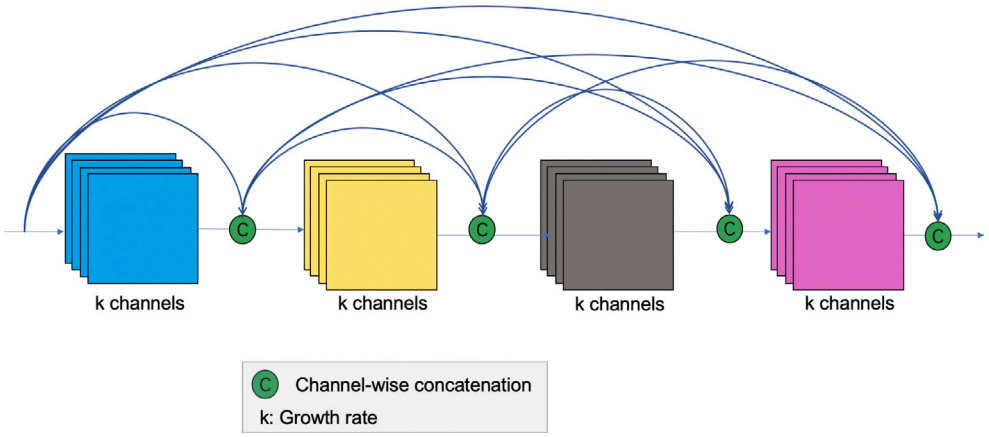

Fig. 2. The general dense block model proposed in this paper. Feature map sizes match within each block.

Table 1
D-SegNet architecture for plant segmentation. The growth rate of k for the whole network is 32. Note that each Up/Down sampling block shown in the table corresponds to the sequence Conv+BN+ReLU+Pooling/Upsampling.

Layers	Output size	D-SegNet ($k = 32$)
Input layer	224 × 224	
Downsample Block (1)	112 × 112	3 × 3 *conv*, stride 1, 2 × 2 max pool, stride 2
Downsample Block (2)	56 × 56	3 × 3 *conv*, stride 1, 2 × 2 max pool, stride 2
Dense Block (1)	56 × 56	$\begin{bmatrix} 1\times1\ conv \\ 3\times3\ conv \end{bmatrix} \times 12$
Downsample Block (3)	28 × 28	3 × 3 *conv*, stride 1, 2 × 2 max pool, stride 2
Dense Block (2)	28 × 28	$\begin{bmatrix} 1\times1\ conv \\ 3\times3\ conv \end{bmatrix} \times 24$
Downsample Block (4)	14 × 14	3 × 3 *conv*, stride 1, 2 × 2 max pool, stride 2
Dense Block (3)	14 × 14	$\begin{bmatrix} 1\times1\ conv \\ 3\times3\ conv \end{bmatrix} \times 16$
Downsample Block (5)	7 × 7	3 × 3 *conv*, stride 1, 2 × 2 max pool, stride 2
Upsample Block (1)	14 × 14	3 × 3 *conv*, stride 1, 2 × 2 max pool, stride 2
Dense Block (4)	14 × 14	$\begin{bmatrix} 1\times1\ conv \\ 3\times3\ conv \end{bmatrix} \times 16$
Upsample Block (2)	28 × 28	3 × 3 *conv*, stride 1, 2 × 2 max pool, stride 2
Dense Block (5)	28 × 28	$\begin{bmatrix} 1\times1\ conv \\ 3\times3\ conv \end{bmatrix} \times 8$
Upsample Block (3)	56 × 56	3 × 3 *conv*, stride 1, 2 × 2 max pool, stride 2
Dense Block (6)	56 × 56	$\begin{bmatrix} 1\times1\ conv \\ 3\times3\ conv \end{bmatrix} \times 4$
Upsample Block (4)	112 × 112	3 × 3 *conv*, stride 1, 2 × 2 max pool, stride 2
Upsample Block (5)	224 × 224	3 × 3 *conv*, stride 1, 2 × 2 max pool, stride 2
Output layer	224 × 224	

Therefore, the D-SegNet is a lightweight network architecture (see Table 1). The number of feature maps of an output remains $4 \cdot k$, which is the same as in [28]. Note that the feature map size only changes in the Downsample/Upsample Blocks. Thus, the feature map size remains the same in all dense blocks, which contains large spatial information in small feature maps.

Different from other semantic segmentation architectures, the D-SegNet encourages feature reuse and prevents gradient vanishing problems. The main reason is that the D-SegNet is a lightweight network, which benefits from dense blocks. Its network structure is narrow, and only needs a few parameters. The number of feature maps of each convolutional layer output in the dense block is very small, instead of hundreds of thousands of outputs like in other networks. Within the dense block, each layer has direct access to the gradients from the loss function and the original input signal. On the other hand, a dense connection is equivalent to directly connecting an input and loss at each layer. Thus, it can mitigate the phenomenon of gradient vanishing when the depth of the network is deep.

2.4. Loss functions used for D-SegNet training

This section discusses the loss functions used to train the D-SegNet architectures. The output from the proposed architecture is a four-dimensional (4D) tensor $Z \in \mathbb{R}^{N,C,H,W}$, where N denotes the batch size of the output tensor, C represents the number of channels (or depth), H is the height and W the tensor width. Colour images have three channels (i.e. $C = 3$) for each channel red, green, and blue, also known as the red, green, blue (RGB) representation. For brevity, Z is considered to be a two-dimensional (2D) matrix of size N, C, containing elements $z_{n,c}$, which is a 2D matrix of size H, W. Prior to computing the cross-entropy loss, the segmented label map output Z needs to be converted to probabilities through the softmax function

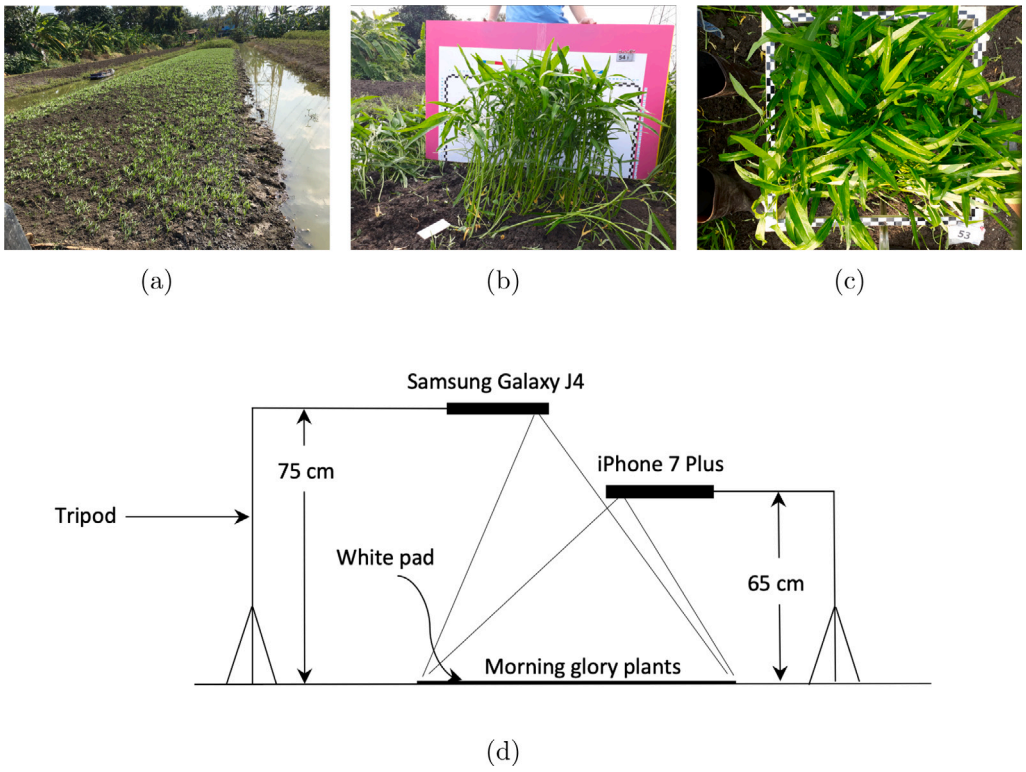

Fig. 3. Environment and construction of image acquisition. (a) Plant culture environment; (b) and (c) Mature plant; (d) Side view of image acquisition construction.

$\phi_{soft}(z_{n,c})$ in the compression layer as follows

$$\phi_{soft}(z_{n,c}) = \frac{\exp(z_{n,c})}{\sum_{i=1}^{C} \exp(z_{n,i})}. \tag{2}$$

The threshold value of the softmax function is set up to 0.6. The output of the softmax layer is then used to compute the cross-entropy loss $\mathcal{L}_{cross}$ which can be calculated from

$$\mathcal{L}_{cross}(z_{n,c}, \hat{y}_{n,c}) = -\sum_{c=1}^{C} \hat{y}_{n,c} \log(\phi_{soft}(z_{n,c})). \tag{3}$$

The softmax output is computed by considering the exponent of the output $z_{n,c}$ from the final layer. This is divided by the sum of the exponential outputs across the channel dimension C. Then, considering the outputs from the softmax layer $\phi_{soft}(z_{n,c})$, the respective target label for that output $\hat{y}_{n,c}$ is used to compute the cross-entropy loss. Here, the term *log* refers to the natural logarithm of the softmax layer element.

The cross-entropy loss is averaged across all N mini-batches of segmented label outputs and target labels, e.g. $\sum_{n=1}^{N} \mathcal{L}_{cross}$. The averaged loss is then backpropagated for updating the layers of the D-SegNet end-to-end using the chain rule and the update rule. The update rule sums the old weights and biases of a specific layer with the differential of the cross-entropy error with respect to the weights and biases of all the convolutional layers multiplied by the learning rate. The learning rate set during training of the D-SegNet is 0.1.

3. Experimental data and data pre-processing

This section describes the data used to evaluate the D-SegNet architecture and the data pre-processing steps. In order to evaluate the performance of D-SegNet, the image data needs to be annotated, which is a complex and time-consuming process, involving the training–validation data. To annotate the data, a semi-automatic labelling method was used consisting of Faster Region-based Convolutional Neural Network (Faster R-CNN) [35] and K-Means clustering [10], which are described specifically in the following Section 3.1.2. This approach improves the efficiency and accuracy of annotation. The original and labelled dataset are available on GitHub [30,31] .

Table 2
The specification of the data collection devices and their distances from the target.

Device	Height	Illuminance	Rear camera	Aperture
Samsung Galaxy J4	65 cm	620-680 LUX	13 MP	$f/1.9$
iPhone 7 plus	75 cm	620-680 LUX	12 MP	$f/1.8$

3.1. Morning glory plant images

A total of 2018 images of the morning glory plant were collected during different phases of the plant growth. These images and the developed deep learning approaches are aimed at helping farmers to decide when to harvest the plant, at the period when the plant is with the desired size and weight. In each of the images, there are 4 or 5 morning glory plants. The plants were laid on white space with tick marks on three edges to indicate the size of the plants. Each side tick mark is 50 centimetres long. The plants in each image were weighed and the total weight was recorded (in grams). The hardware specifications used for data collection and their relevant distances from the target (height) are summarized in Table 2.

Each image was taken by two smartphones, a Samsung Galaxy J4 and an iPhone 7 plus. The Samsung Galaxy J4 was installed at 65 centimetres above the plants, while the iPhone 7 Plus was installed at 75 centimetres above the plants. The images are of size 4032×3024 or 4128×3096. The Samsung Galaxy J4 mobile phone has a 13 MP rear camera with $f/1.9$ aperture, and the iPhone 7 Plus has a 12 MP rear camera with $f/1.8$ aperture. Fig. 3 shows the plant environment and image acquisition device. All images were taken in a laboratory with an illuminance of approximately 620–680 LUX. The light was mainly from the fluorescent lightbulbs installed in the laboratory with some effect from natural light through the windows. There were only a few images in that the illuminance that fell below or above the range mentioned.

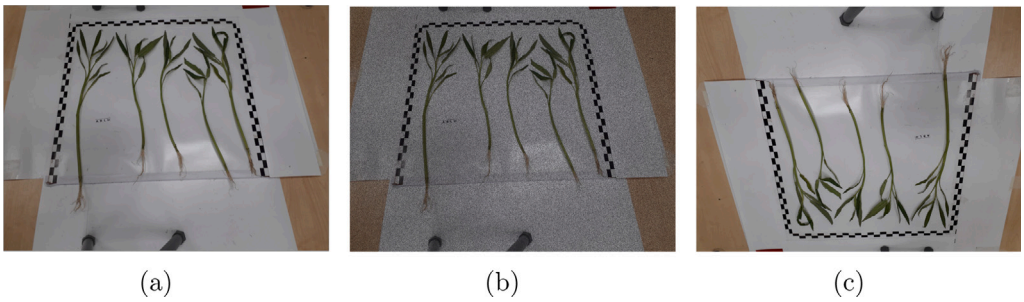

Fig. 4. Morning glory plant images: (a) original image, (b) image with added noise and (c) image with a different illumination background. Images (b) and (c) are obtained as a result of the data augmentation.

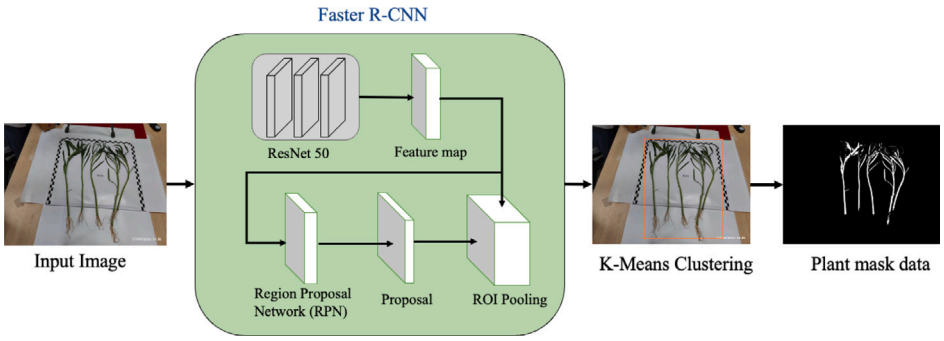

Fig. 5. Image data annotation steps.

This could perhaps be due to the sensitivity of the device or the shadow from the experimenters.

The experiments of automatic plant segmentation are conducted on this dataset. The dataset was collected by the team from King Mongkut's University of Technology Thonburi, at the geographical location with latitude and longitude (13.39 °N, 100.29 °E) in Thailand.

3.1.1. Image pre-processing

Data augmentation has an important role in deep learning [36] effectively improves the model robustness and generalization capacity.

Generally, data augmentation includes the generation of additional images based on random flips, flexible rotation and different illumination. Fig. 4 shows the augmented data. Figs. 4(a) and 4(b) have different noise and illumination backgrounds, respectively. Fig. 4(c) is obtained after a rotation of the original image at 180°. These operations expand the original data set from 2018 to 6054 images, which will help to train and evaluate the segmentation networks.

3.1.2. Image data annotation

The overlap and thickness of leaves make the manual annotation difficult and create challenges to the training of convolutional neural networks. Thus, a semi-automatic annotation method is applied, whose main steps are summarized in Fig. 5.

The Faster R-CNN algorithm [35] is applied to detect the plant. Faster R-CNN is a state-of-the-art object detection network, which consists of region-based proposal algorithms able to find the object location [37]. R-CNN is composed of a feature extraction network followed by two networks. The feature extraction network is a pre-trained CNN that forms the feature map. The first network ResNet-50 [33] is the feature extraction network of the detector. The second network is a region proposal network (RPN), which generates object proposals. The RPN decides the positive or negative of anchors by the softmax function. The bounding box regression is used to fix anchors and get precise accuracy.

The region of interesting (ROI) pooling layer processes features maps and proposals for further prediction. The area outside of the region box is set up as a black background. Next, the k-means clustering algorithm [10,38] is applied to annotate the leaves within the bounding box area. The k-means algorithm clusters similar data points into k groups. The target number k refers to the number of centres in the cluster. Every data point is allocated to the nearest cluster while keeping the centroids as small as possible. In this paper, a fixed number k is set equal to 2.

To avoid complex data processing [39], the annotated data consider leaf pixels as white and the other as black, as shown in Fig. 6. It is saved in a PNG format with high resolution 4128 × 3096. The high-resolution image is stored but it is not processed by the convolutional neural network since such an image would require a large amount of memory space. This would require significant computational power from the GPU. Therefore, the size of original and annotated images is processed as 224 × 224 × 3 and 224 × 224.

3.2. Benchmark dataset

A popular public dataset from ImageCLEF 2021 (Pl@ntLeaves), [32] is used to evaluate our proposed segmentation method. This dataset consists of leaf herbarium and comprises 1956 images, from which 1190 images are used for training, and 256 for testing. Sample leaf images are shown in Fig. 7.

3.3. K-fold cross validation

Cross-validation is a resampling method that is used to evaluate models on a limited data sample [40]. In K-fold cross-validation, the entire dataset is split into K groups randomly. For every fold, one out of K subsets is chosen as the validation set and K-1 subsets are used for training. It helps to avoid the overfitting problem and to improve model performance with the small dataset. The diagram of K-fold cross-validation is shown in Fig. 8.

Experiments are conducted for 10-fold cross-validation with the morning glory plant dataset and ImageCLEF dataset. Each dataset is divided into 10 subsets stochastically. For each cross-validation iteration, the model is trained using 9 subsets and is tested to the other

(a) (b)

Fig. 6. Images of the morning glory plant. (a) Original image; (b) Annotated image.

(a) (b)

Fig. 7. ImageCLEF dataset: a herbarium sheet.

subset as the validation set. The procedure is then repeated for 10-folder cross-validation. Each subsample is used once as a validation subset.

4. Results and discussion

4.1. Evaluation metrics

This section presents the evaluation metrics used for the performance validation of the D-SegNet algorithm. These include precision, recall, and the F1 score. The precision–recall metrics [41] allow further evaluation for classification beyond simple accuracy measures that do not take into account the problem of class imbalance.

First, the number of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) samples are calculated. The TP are predictions that are classified correctly for the inspected class. The TN are other predictions that are correct but for negative samples (i.e. not inspected). Then, false positives are miss-classified negative samples and the true negatives are negatives samples that are correctly classified. The precision and recall values are computed as shown in Eqs. (4) and (6). In addition, the IoU is also an important metric and it is used to evaluate the algorithms' performance.

The precision metric [42] is used to compute the number of correctly predicted positive categories of a classification system

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{4}$$

Precision is defined as the ratio of the total number of TPs divided by the sum of all TPs and FPs. The average of per-class precision [43] is

$$\text{Precision}_M = \frac{1}{l} \sum_{i=1}^{l} \text{Precision}, \tag{5}$$

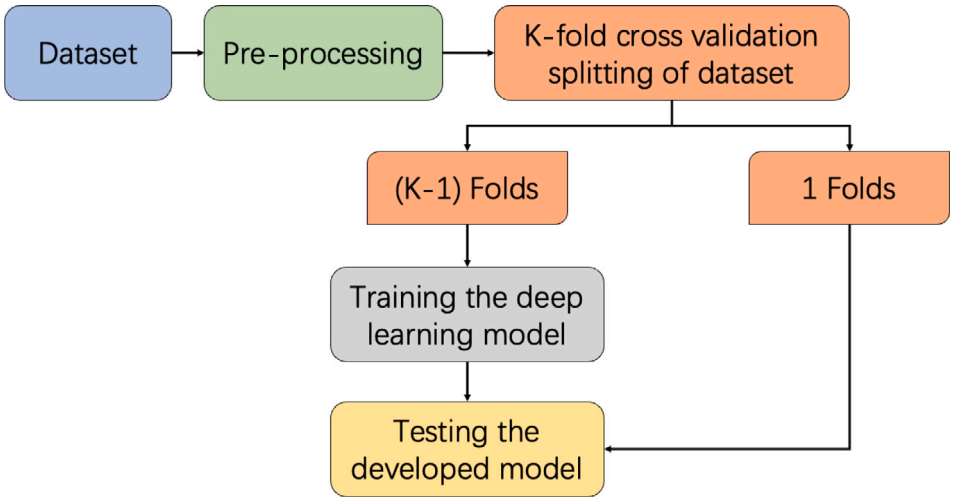

Fig. 8. The process of K-fold cross-validation in training the deep learning model.

which is the arithmetic mean of all the summary precision values by the number of classes l.

The recall metric

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{6}$$

is used to define the number of correct positive predictions that are achievable from all of the positive predictions. The average per-class recall is identified as

$$\text{Recall}_M = \frac{1}{l}\sum_{i=1}^{l} \text{Recall}. \tag{7}$$

The Recall_M [43] focuses on the per-class effectiveness of class labels. A good model expects to get high values on both precision and recall. However, it is difficult to decide the model performance when the precision and recall are reaching different extremums. Thus, it is necessary to use the F1-score in the evaluation. The F1-score_M [43], on the other hand, summarizes the Precision_M and the recall of a classifier system into a single metric

$$\text{F1-score}_M = 2 \times \frac{(\text{Precision}_M \times \text{Recall}_M)}{(\text{Precision}_M + \text{Recall}_M)}. \tag{8}$$

The model evaluation should not only use the statistic metrics but also need a qualitative evaluation from the actual segmentation view. The intersection over union (IoU) [44] as an important evaluation index in semantic segmentation measures the overlap of the ground truth and prediction region. It is also calculated:

$$\text{IoU} = \frac{\text{Ground Truth} \bigcap \text{Prediction}}{\text{Ground Truth} \bigcup \text{Prediction}}, \tag{9}$$

where $\bigcap$ denotes the intersection operator and $\bigcup$ represents the union operator. The IoU is generally calculated based on categories, which is to accumulate the IoU value of each category. The IoU value is to average the sum IoU results of each category to obtain a global evaluation. Therefore, the IoU is actually the mean value, that is, the average crossover ratio (mean IoU).

In the next section, the performance of the segmented results is discussed in detail while comparing different architectures.

4.2. Experimental results

In order to evaluate the performance of the D-SegNet algorithm, we compare it with traditional edge-based segmentation and with the standard SegNet algorithm.

From Fig. 9, we can see that D-SegNet provides quite clear contours of the segmented plant, whereas the classical Sobel edge-based segmentation [45] visually gives less accurate results. Morphological operations [46,47] are used to connect the edge pixels into meaningful edges.

Table 3
Performance of different segmentation methods in morning glory plant.

Metrics	Based on edge detection	SegNet-Basic	SegNet	D-SegNet
Precision_M	67.38%	96.57%	97.48%	**98.31%**
Recall_M	62.25%	88.74%	89.64%	**90.92%**
F1-score_M	64.71%	92.48%	93.39%	**94.47%**

Table 4
Performance of different segmentation methods in ImageCLEF (Pl@ntleaves) dataset.

Metrics	Based on edge detection	SegNet-Basic	SegNet	D-SegNet
Precision_M	46.35%	90.98%	93.79%	**96.57%**
Recall_M	41.68%	84.87%	87.03%	**88.54%**
F1-score_M	43.89%	87.81%	90.28%	**92.38%**

There are two steps in morphological operations: erosion and dilation. Dilation expands the image boundaries to make sure the boundary pixels are connected. Erosion shrinks the image pixels slightly, which removes noise pixels from the object boundary. Finally, the object is segmented from the image. However, it cannot provide a coherent and precise leaf shape. Although this does not mean traditional methods become obsolete, D-SegNet indeed solves complex segmentation problems with super-human accuracy. The big data training and plentiful feature computation provide descriptive and salient features to predict the underlying patterns. Thereby D-SegNet get a better performance than the traditional computer vision method.

In this paper, the proposed approach, SegNet-Basic [48] and SegNet [49] methods are trained. In order to provide a fair comparison of the performance of the considered deep learning architectures, the same GPU and environment system configurations are applied to all of them. A 10-fold cross-validation strategy is applied to each data set. Then 9 folds are used for training, and the other one fold is used for testing. The procedure is repeated 10 times to make sure each fold has been used for testing once.

Fig. 10 displays a group of segmented images. Based on subjective observation, the two framed parts in each figure show different segmentation results. SegNet-Basic and SegNet are hard to segment the curve leaf, while D-SegNet gives a clear shape. It is expected that the proposed approach can capture many features through the dense block. Obviously, the D-SegNet achieves higher performance than the other algorithms. The SegNet-Basic [48] is a more lightweight network than SegNet, which comprises 4 pairs of encoder–decoders. As shown in Fig. 10(b), SegNet-Basic can take the coarse contour. Several pixels

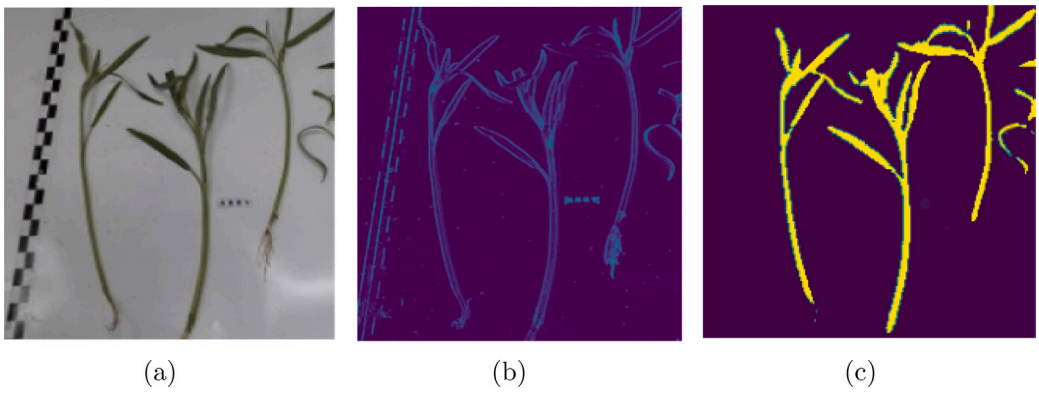

Fig. 9. Segmentation results. (a) Original image; (b) Segmentation based on edge detection with Sobel operator; (c) D-SegNet.

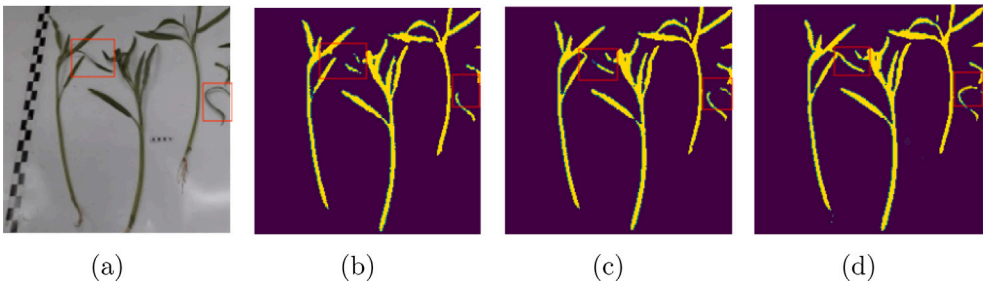

Fig. 10. Results of transitional steps in proposed model for automating the segmentation process. (a) Original image; (b) SegNet-Basic; (c) SegNet; (d) D-SegNet.

of leaf edge have been lost. In addition, SegNet as the state-of-the-art algorithm [49] is selected for comparison. It is a fully convolutional network with a pixel-level classification network that has 5 pairs of encoder–decoders. SegNet segments well the plant from the lab background compared with SegNet-Basic as seen in Fig. 10(c). However, when SegNet compares with the D-SegNet segmentation method in this paper, it can be found that the SegNet segmentation method loses some details, as shown in Fig. 10(c). The SegNet method is sensitive to the close areas between plants, leaves overlapping and interference with light conditions. Thus, SegNet loses some features on leaves, and it cannot segment the special position of the plant in the original image. In this paper, the D-Segnet method solved these problems, which can extract plant boundary contour from the complex background accurately.

Tables 3 and 5 list average quantitative segmentation results calculated over 10-fold cross-validation on the morning glory plant dataset. The numerical results denote precision, recall, F1-score and IoU in the four segmentation methods. There is a significant difference between segmentation based on edge detection and other deep learning algorithms. All mean metric values of deep learning algorithms are higher than that of the traditional image processing algorithm. The performance of D-SegNet outperforms that of SegNet-Basic and SegNet with respect to both visual and numerical results. Both SegNet-Basic and SegNet are encoder–decoder networks. The size of the feature map is going small after each convolution. The common problem is the convolution operation between layers may lose important features. The D-SegNet is designed to break this bottleneck. This is thanks to the concatenation between layers of the dense block in the D-SegNet. It makes the feature extracted from the encoder–decoder, which directly connects to each layer in the dense block. The D-Segnet will learn more features than the SegNet-Basic and SegNet, which makes D-SegNet lead the performance competition. The performance differences are obvious from Table 4. It shows the mean values of precision, recall and F1-score with four different algorithms on the ImageCLEF dataset. The D-SegNet gives accurate segmentation results on both datasets.

Table 5
Comparison of different segmentation methods in performance.

Network	mean IoU
Based on edge detection	56.37%
SegNet-Basic	90.56%
SegNet	88.58%
D-SegNet	**90.64%**

4.3. Additional considerations

The phone images and the pixels of leaves might be not accurate enough. For each group, we arrange the plants 4 times and take 4 images. However, the leaf of the plant grows discretely and is displayed on the board randomly. Thus, the same group of plants could have different labelling pixels. Moreover, the pixels from overlapping leaves are counted only once. The relationship between plant pixels and weight is an important point in automatic yield estimation. Therefore, there are two methods to solve this problem. It is difficult to label the overlapped leaves pixel by pixel. Secondly, the sample of each group plant is increased. For the same group of plants, there are a lot of images from different plant positions.

In addition, the proposed D-SegNet algorithm segments only the plant in the mature stage and demonstrates a better performance than the state-of-the-art algorithm [49]. However, it did not fully grasp the plant in each growing node. The timing of harvesting still relies on labour. With the growth of the morning glory plant, the physical shape is changing dynamically. In the future, we would like to test the different morphological shapes, as it is important to understand the morphological and physiological characteristics of the morning glory plant.

All images are taken by a smartphone. It demonstrates the potential of using a smartphone which is becoming ubiquitous and affordable to many farmers. For future applications, images taken by the farmers may be sent to an automated system to predict the yield. The techniques tested in this study show that noisy images could be processed and used

to predict yield with reasonably good accuracy. This is the first step toward developing an automated system for yield prediction for smart-farm planning and monitoring. In addition, the D-SegNet algorithm is not only applied to automatic yield prediction in morphology but also contributes to weed detection, for disease detection and distinction, as well as other related tasks. For robust comparison, the evaluation metrics are averaged over 10-fold cross-validation. The results and evaluation techniques suggest that the proposed approaches are not only superior with regard to segmentation capabilities but also with respect to reliability.

5. Conclusions

A highly accurate deep learning-based segmentation method called D-SegNet, is proposed in this paper, which can be used in automated yield prediction and precision agriculture. Compared to traditional non-deep learning methods, D-SegNet achieves pixel-level segmentation. The dense block structure in D-SegNet enhances the feature propagation to outperform state-of-the-art deep learning algorithms such as SegNet. The concatenations between feature maps can better extract object information and can be used to input sequential data. The trained network can accurately detect segmentation details such as very small overlapped leaves. Experiments were carried out and four metrics were used to compare our method with other competing methods. The results show that D-SegNet achieved 0.9820, 0.9064 and 0.9426 in precision, recall and F1-score, respectively. The IoU calculated over 2421 untrained plant tasks was 0.9056. These results demonstrate that D-SegNet can segment the plant accurately and improve the general performance of the network compared to prior methods.

The morning glory plant is a popular crop in the agriculture business and its yield prediction is subject to exciting research. This research could help small or community-based farms better to estimate the amount or yield of their produce. With an accurate yield estimation, the buyer of the product can save costs through better planning its sourcing strategy and simultaneously better serve the needs of the customers. This further promotes the business of the buyer, leading to closer collaboration between the farmers and the buyer.

In summary, this new plant segmentation network, D-SegNet, could form an important component of future cloud-based machine learning systems to predict crop yield from noisy smartphone images taken in the field.

CRediT authorship contribution statement

Jingxuan Su: Conceptualization, Methodology, Software, Validation, Writing – original draft, Writing – review & editing, Visualization. **Sean Anderson:** Conceptualization, Methodology, Validation, Formal analysis, Writing – review & editing. **Mahed Javed:** Conceptualization, Methodology, Validation, Formal analysis, Writing – review & editing. **Charoenchai Khompatraporn:** Dataset, Conceptualization, Methodology, Validation, Formal analysis, Writing – review & editing. **Apinanthana Udomsakdigool:** Dataset, Conceptualization, Methodology, Validation, Formal analysis, Writing – review & editing. **Lyudmila Mihaylova:** Conceptualization, Methodology, Validation, Formal analysis, Writing – review & editing.

Declaration of competing interest

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

Data availability

The data that has been used is confidential.

Acknowledgements

This work is supported by the "Smart Farm Planning and Monitoring System" Project funded by the UK Royal Academy of Engineering under the Industry Academia Partnership Programme (IAPP) and the Thailand Integrated Research Program [grant number IAPP18-19299] and the Thailand Integrated Research Program [grant number 47330]. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

References

[1] R. Gebbers, V.I. Adamchuk, Precision agriculture and food security, Science 327 (5967) (2010) 828–831.
[2] A.C. Tyagi, Towards a second green revolution, Irrigation Drainage 65 (4) (2016) 388–389.
[3] M.H. Anisi, G. Abdul-Salaam, A.H. Abdullah, A survey of wireless sensor network approaches and their energy consumption for monitoring farm fields in precision agriculture, Precis. Agricult. 16 (2) (2015) 216–238.
[4] A. Savla, N. Israni, P. Dhawan, A. Mandholia, H. Bhadada, S. Bhardwaj, Survey of classification algorithms for formulating yield prediction accuracy in precision agriculture, in: Proc. of the International Conference on Innovations in Information, Embedded and Communication Systems, ICIIECS, IEEE, 2015, pp. 1–7.
[5] A. Nigam, S. Garg, A. Agrawal, P. Agrawal, Crop yield prediction using machine learning algorithms, in: Proceedings of the Fifth International Conference on Image Information Processing, ICIIP, IEEE, 2019, pp. 125–130.
[6] S.A. Tsaftaris, M. Minervini, H. Scharr, Machine learning for plant phenotyping needs image processing, Trends Plant Sci. 21 (12) (2016) 989–991.
[7] Z. Wang, K. Wang, F. Yang, S. Pan, Y. Han, Image segmentation of overlapping leaves based on Chan–Vese model and Sobel operator, Inf. Process. Agric. 5 (1) (2018) 1–10.
[8] G.J. Brindha, E. Gopi, An hierarchical approach for automatic segmentation of leaf images with similar background using kernel smoothing based Gaussian process regression, Ecol. Inform. 62 (2021) 101323.
[9] A. Aggelopoulou, D. Bochtis, S. Fountas, K.C. Swain, T. Gemtos, G. Nanos, Yield prediction in apple orchards based on image processing, Precis. Agricult. 12 (3) (2011) 448–456.
[10] T. Kanungo, D.M. Mount, N.S. Netanyahu, C.D. Piatko, R. Silverman, A.Y. Wu, An efficient k-means clustering algorithm: Analysis and implementation, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 881–892.
[11] S. Dargan, M. Kumar, M.R. Ayyagari, G. Kumar, A survey of deep learning and its applications: a new paradigm to machine learning, Arch. Comput. Methods Eng. 27 (4) (2020) 1071–1092.
[12] L. Saxena, L. Armstrong, A survey of image processing techniques for agriculture, in: Proceedings of Asian Federation for Information Technology in Agriculture, Australian Society of Information and Communication Technologies in Agriculture, 2014, pp. 401–413.
[13] J.M. Guerrero, G. Pajares, M. Montalvo, J. Romeo, M. Guijarro, Support vector machines for crop/weeds identification in maize fields, Expert Syst. Appl. 39 (12) (2012) 11149–11155.
[14] M.A. Sartin, A.C. Da Silva, C. Kappes, Image segmentation with artificial neural network for nutrient deficiency in cotton crop, J. Comput. Sci. (2014) 1084–1093.
[15] N. Kussul, M. Lavreniuk, S. Skakun, A. Shelestov, Deep learning classification of land cover and crop types using remote sensing data, IEEE Geosci. Remote Sens. Lett. 14 (5) (2017) 778–782.
[16] J. Naranjo-Torres, M. Mora, R. Hernández-García, R.J. Barrientos, C. Fredes, A. Valenzuela, A review of convolutional neural network applied to fruit image processing, Appl. Sci. 10 (10) (2020) 3443.
[17] D. Komura, S. Ishikawa, Machine learning methods for histopathological image analysis, Comput. Struct. Biotechnol. J. 16 (2018) 34–42.
[18] P. Bosilj, E. Aptoula, T. Duckett, G. Cielniak, Transfer learning between crop types for semantic segmentation of crops versus weeds in precision agriculture, J. Field Robotics 37 (1) (2020) 7–19.
[19] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, et al., Recent advances in convolutional neural networks, Pattern Recognit. 77 (2018) 354–377.
[20] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.
[21] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[22] S.G. Sodjinou, V. Mohammadi, A.T.S. Mahama, P. Gouton, A deep semantic segmentation-based algorithm to segment crops and weeds in agronomic color images, in: Information Processing in Agriculture, Elsevier, 2021.
[23] S. Zagoruyko, N. Komodakis, Wide residual networks, 2016, arXiv preprint arXiv:1605.07146.
[24] Q. Zhang, Y. Liu, C. Gong, Y. Chen, H. Yu, Applications of deep learning for dense scenes analysis in agriculture: A review, Sensors 20 (5) (2020) 1520.

[25] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, P. Martinez-Gonzalez, J. Garcia-Rodriguez, A survey on deep learning techniques for image and video semantic segmentation, Appl. Soft Comput. 70 (2018) 41–65.

[26] A. Singh, B. Ganapathysubramanian, A.K. Singh, S. Sarkar, Machine learning for high-throughput stress phenotyping in plants, Trends Plant Sci. 21 (2) (2016) 110–124.

[27] K. Cho, B. Van Merriënboer, D. Bahdanau, Y. Bengio, On the properties of neural machine translation: Encoder-decoder approaches, 2014, arXiv preprint arXiv:1409.1259.

[28] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.

[29] L. Bottou, Large-scale machine learning with stochastic gradient descent, in: Proceedings of COMPSTAT'2010, Springer, 2010, pp. 177–186.

[30] S. Kanokwan, C. Khompatraporn, Morning glory plant datset, 2021, https://github.com/S-KANOKWAN?tab=repositories. Website.

[31] C. Kittipong, C. Khompatraporn, Morning glory plant datsets, 2021, https://github.com/CH-KITTIPONG?tab=repositories. Website.

[32] H. Goëau, P. Bonnet, A. Joly, Overview of PlantCLEF 2021: cross-domain plant identification, in: Working Notes of CLEF 2021-Conference and Labs of the Evaluation Forum, Vol. 2936, 2021, pp. 1422–1436.

[33] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[34] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: Proceedings of the International Conference on Machine Learning, PMLR, 2015, pp. 448–456.

[35] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, Adv. Neural Inf. Process. Syst. 28 (2015) 91–99.

[36] A. Dorward, E. Chirwa, A Review of Methods for Estimating Yield and Production Impacts, Technical report, Centre for Development, Environment and Policy, SOAS, University of London, UK, 2010.

[37] C. Feng, D. Zhao, M. Huang, Image segmentation and bias correction using local inhomogeneous intensity clustering (LINC): a region-based level set method, Neurocomputing 219 (2017) 107–129.

[38] H. Lu, Q. Gao, X. Zhang, W. Xia, A multi-view clustering framework via integrating K-means and graph-cut, Neurocomputing 501 (2022) 609–617.

[39] L. Ma, M. Wang, J. Dong, K. Peng, A novel distributed detection framework for quality-related faults in industrial plant-wide processes, Neurocomputing 492 (2022) 126–136.

[40] P. Refaeilzadeh, L. Tang, H. Liu, Cross-validation, Encycl. Database Syst. 5 (2009) 532–538.

[41] C. Goutte, E. Gaussier, A probabilistic interpretation of precision, recall and F-score, with implication for evaluation, in: Proceedings of the European Conference on Information Retrieval, Springer, 2005, pp. 345–359.

[42] M. Everingham, S. Eslami, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes challenge: A retrospective, Int. J. Comput. Vis. 111 (1) (2015) 98–136.

[43] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Inf. Process. Manage. 45 (4) (2009) 427–437.

[44] M.A. Rahman, Y. Wang, Optimizing intersection-over-union in deep neural networks for image segmentation, in: Proceedings of the International Symposium on Visual Computing, Springer, 2016, pp. 234–244.

[45] W. Gao, X. Zhang, L. Yang, H. Liu, An improved Sobel edge detection, in: Proc. of the 3rd International Conference on Computer Science and Information Technology, Vol. 5, IEEE, 2010, pp. 67–71.

[46] F. Meyer, S. Beucher, Morphological segmentation, J. Vis. Commun. Image Represent. 1 (1) (1990) 21–46.

[47] M. Pesaresi, J.A. Benediktsson, A new approach for the morphological segmentation of high-resolution satellite imagery, IEEE Trans. Geosci. Remote Sens. 39 (2) (2001) 309–320.

[48] V. Badrinarayanan, A. Handa, R. Cipolla, SegNet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling, 2015, arXiv preprint arXiv:1505.07293.

[49] V. Badrinarayanan, A. Kendall, R. Cipolla, SegNet: A deep convolutional encoder-decoder architecture for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (12) (2017) 2481–2495.

Jingxuan Su obtained a B.Eng. degree in Electronic Information Engineering in 2018 from the Shenyang University of Technology in China. She received her M.Sc. degree in Control, Systems and Signal Processing at the University of Southampton, UK, in 2019. She is currently a Ph.D. student in the Department of Automatic Control and Systems Engineering, at the University of Sheffield, UK. Her research interests include machine learning and deep learning methods in the smart farm, image segmentation, object detection and data imbalance problems.

Sean R. Anderson received the M.Eng. degree in control systems engineering and the Ph.D. degree in nonlinear system identification and predictive control for fast-sampled systems from the Department of Automatic Control and Systems Engineering, The University of Sheffield, Sheffield, U.K., in 2001 and 2005, respectively. From 2005 to 2010, he was a Research Associate with the Neural Algorithms Research Group, The University of Sheffield. From 2010 to 2011, he was a Research Associate with the Department of Automatic Control and Systems Engineering, The University of Sheffield, where he became a Lecturer in 2012 and has been a Senior Lecturer there since 2015.

Mahed Javed holds a Ph.D. from the Department of Automatic Control and Systems, Sheffield, UK. He obtained his undergraduate degree in Engineering from the University of Liverpool, UK in 2016 and a postgraduate degree from the University of Sheffield, UK in 2017. He worked as a software engineer during his University study period more than once. Firstly, as a software designer for the IMechE Unmanned Aerial Vehicle challenge from 2015–2016. Secondly as a software developer for the Sheffield University Nova Balloon Telescope (SUNBYTE). His research interests include machine learning and deep learning in the specific area of uncertainty quantification and adversarial robustness.

Charoenchai Khompatraporn is an Associate Professor at the Department of Production Engineering, Faculty of Engineering, and the Associate Dean on Academic Affairs of the Graduate School of Management and Innovation, King Mongkut's University of Technology Thonburi, Thailand. He holds a Ph.D. from University of Washington, and a master's degree from Georgia Institute of Technology, USA. He received a bachelor's degree from Rensselaer Polytechnic Institute, New York. His research interests include supply chain and logistics management, applied operational research, optimization algorithms, and industrial sustainable operations. He has been working closely with both public and private sectors such as Thai Red Cross, steel manufacturers, agri-businesses, energy providers and social enterprises on various logistics, operations, and sustainability issues.

Apinanthana Udomsakdigool was an Assistant Professor at the Department of Production Engineering, Faculty of Engineering, King Mongkut's University of Technology Thonburi, Thailand. She received a D.Eng. from Asian Institute of Technology, a master's degree from Chulalongkorn University and a bachelor's degree from King Mongkut's University of Technology Thonburi, Thailand. Her research interests included advanced planning and scheduling, applied operational research in industry, production management and combinatorial optimization model.

Lyudmila Mihaylova is Professor of Signal Processing and Control at the Department of Automatic Control and Systems Engineering at the University of Sheffield, United Kingdom. She received MEng degree in Systems and Control Engineering, M.Sc. degree in Applied Mathematics and Informatics and her Ph.D. degree is in Systems and Control Engineering, all from the Technical University of Sofia, Bulgaria. Her research is in the areas of autonomous systems and machine learning with various applications such as navigation, surveillance, and sensor networks. She is Associate Editor-in-Chief for the IEEE Transactions on Aerospace and Electronic Systems since 2021 and a Subject Area Editor for the Elsevier Signal Processing Journal since 2022. She was the president of the International Society of Information Fusion (ISIF) from 2016 to 2018. She is on the Board of Directors of ISIF. She has been serving for organizing conferences — as the general vice chair of UKCI'2022, a program chair for the International Conference on Information Fusion, Fusion 2022, technical chair for Fusion 2021, publicity chair for IEEE MFI' 2021 and UKCI 2021. She was the general vice-chair for the International Conference on Information Fusion 2018 (Cambridge, UK), of the IET Data Fusion & Target Tracking 2014 and 2012 Conferences, publications chair for ICASSP 2019 (Brighton, UK) and others.