Beet Seedling and Weed Recognition Based on Convolutional Neural Network and Multi-modality Images

Jun Sun (sun2000jun@sina.com)
Jiangsu university https://orcid.org/0000-0003-3019-6086
Yu Yang
Jiangsu University
Xiaofei He
JSU: Jiangsu University
Long Wang
Jiangsu University
Xiaohong Wu
Jiangsu University
Jifeng Shen
Jiangsu University

Research

Keywords: beets and weeds, deep learning, deformable convolution, multi-modality images, object detection

Posted Date: May 25th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-520181/v1

Version of Record: A version of this preprint was published at Multimedia Tools and Applications on December 14th, 2021. See the published version at https://doi.org/10.1007/s11042-021-11764-5.

Beet seedling and weed recognition based on convolutional neural network and multi-modality images

Jun Sun[1], Yu Yang[1], Xiaofei He[2], Long Wang[3], Xiaohong Wu[4], Jifeng Shen[5]
School of electrical and Information Engineering, Jiangsu University, Zhenjiang, China
Correspondence
Jun Sun, School of electrical and Information Engineering, Jiangsu University, Zhenjiang 212000, China.
Email: sun2000jun@sina.com
Tel：13775544650

Abstract

Background: Difficulties in the recognition of beet seedlings and weeds can arise from a complex background in the natural environment and a lack of light at night. In the current study, a novel depth fusion algorithm was proposed based on visible and near-infrared imagery.

Results: Visible (RGB) and near-infrared images were superimposed at the pixel-level via a depth fusion algorithm and were subsequently fused into three-channel multi-modality images in order to characterize the edge details of beets and weeds. Moreover, an improved region-based fully convolutional network (R-FCN) model was applied in order to overcome the geometric modeling restriction of traditional convolutional kernels. More specifically, for the convolutional feature extraction layers, deformable convolution was adopted to replace the traditional convolutional kernel, allowing for the entire network to extract more precise features. In addition, online hard example mining was

introduced to excavate the hard negative samples in the detection process for the retraining of misidentified samples. A total of four models were established via the aforementioned improved methods. Results demonstrate that the average precision of the improved optimal model for beets and weeds were 84.8% and 93.2%, respectively, while the mean average precision was improved to 89.0%.

Conclusion: Compared with the classical R-FCN model, the performance of the optimal model was not only greatly improved, but the parameters were also not significantly expanded. Our study can provide a theoretical basis for the subsequent development of intelligent weed control robots.

Keywords

beets and weeds, deep learning, deformable convolution, multi-modality images, object detection

Backgroud

The presence of weeds in the field can cause great damage to crop seedlings. More specifically, weeds compete with crops for sunlight and nutrients, thus seriously affecting the photosynthesis of seedlings and increasing the spread of diseases and insect pests. Therefore, the removal of weeds is of great significance in order to maintain crop yield[1]. Generally, the weeding methods based on machine vision are performed under good daylight conditions. However, in order to improve production efficiency and reduce the damage of weeds to crop seedlings, continuous operation at night is required to identify

weeds. Traditional weed elimination methods are time-consuming and laborious, and mainly depends on artificial excavation or pesticide spraying. Moreover, it is difficult to identify weeds that are similar to crops[2]. Pesticide residues produced by spraying not only poses a great threat to human health, but can also damage the ecological environment[3]. With the increasing application of precision agriculture, it is particularly important to establish the real-time accurate identification of weeds and crops for the rational application of pesticides, crop yield increases, the reduction of environmental pollution, and the implementation of intelligent weeding[4-6].

At present, most weed detection methods are developed based on machine vision. A classification method for weed classification based on a back propagation (BP) neural network was proposed[7]. Following the fuzzy classification of the features, a genetic algorithm was used to optimize the network for the identification of weeds. A method to identify weeds based on machine vision during the maize seedling stage was designed[8]. After distortion correction, HSI (Hue, Saturation and Intensity) color space conversion and threshold segmentation, the collected images of maize plants and weeds were identified according to the shape and color features. A weed segmentation network based on an artificial neural network was designed[9]. The single-stage wavelet transform was used to extract weed texture features with 14 texture features selected to optimize the algorithm by using the principal component analysis method. Finally, the features were sent into the neural network in order to identify the weeds. The aforementioned studies combine shallow feature extraction and pattern recognition to identify weeds. However,

65 the feature extraction of such methods is time consuming and the applicability is weak.
66 In addition, due to the influence of the complex field background, the weed characteristics
67 extracted by humans can be ambiguous and uncertain, which consequently results in weed
68 identification based on traditional machine vision at low accuracies.

69 Traditional machine vision methods can only generate low-level image features, which
70 limits the expression ability. However, the convolutional neural network (CNN) has a
71 strong capability to represent image feature without manual feature selection. It can
72 further extract high-level abstract features in images with a high accuracy and efficiency
73 based on the gradient descent and back propagation algorithm[10]. Therefore, CNNs are
74 widely used in image classification[11], object detection[12] and semantic segmentation[13]. The
75 neurons in convolutional layers can extract the primary visual features of the image by
76 local receptive fields, and reduce the number of parameters through weight sharing. The
77 pooling layer can realize the invariance of displacement, scaling and distortion, while
78 simultaneously performing feature dimensionality reduction.

79 Recently, many networks based on pre-trained CNNs have achieved promising results
80 in weed and crop seedling detection. A weed identification method based on a deep CNN
81 and hash code was proposed[14], which was able to effectively compress the high-
82 dimensional features of the weed through a binary hash layer to detect weeds. Andrea
83 used a classification method based on a CNN that identified maize seedlings and weeds
84 by optimizing the number of convolutional kernels on the basis of the original
85 classification network[15]. Joseph Proposed a UAV for weed detection and fixed-point

removal, which limited the use of pesticides and improved the efficiency of weed removal[16]. Results from the aforementioned literature demonstrate that CNNs can not only automatically extract the shallow features (texture, color, etc.) of weeds and crops, but can also learn deeper abstract features. Moreover, CNNs are able to reduce the cost of feature extraction and are more robust for weed detection in a complex environment. Therefore, CNNs have the potential to be applied to detect beet seedlings and weeds. However, the models in the current literature are not sensitive to the feature information of weeds due to the weak light at night and the use of traditional convolutional kernels. This has resulted in feature extraction difficulties and low recognition accuracies.

At present, the computer vision-based detection of crops and weeds in the field is limited. This is a result of the noise commonly present in images of certain scenes (e.g. at night with insufficient light). Thus, the detection process is easily affected changes in light, thus misinterpreting the shape, color, texture and additional information of the detected object, resulting in a poor visual performance. Moreover, traditional CNNs only use the original regular convolutional kernel to extract the features of crops and weeds, with a limited ability in geometric modeling. This may weaken the feature extraction ability of the model. Furthermore, during the training of the model, the uneven distribution of positive and negative samples will also lead to the poor generalization of the model. In view of the above limitations of the current methods, three improved methods were proposed in this study:

(1) A deep fusion algorithm was adopted to fuse the RGB (Red, Green and Blue) and

near-infrared (NIR) images of beets and weeds into three-channel multi-modality images, then the fusion images were sent into CNN for training.

(2) In the feature extraction layer, the traditional convolutional kernel was replaced by the deformable convolution.

(3) The hard negative samples were fully excavated by using online hard example mining (OHEM) in the detection process, and then they were sent into the network for re-identification.

Results

In the current study, an improved R-FCN model was proposed to detect and identify beet seedlings and weeds under poor light (at night) and complex field backgrounds. Based on the classic R-FCN network, the visible and near-infrared images of beets and weeds were fused into three-channel multi-modality images at pixel-level using a deep fusion algorithm. The fusion images were then sent to the convolutional neural network for training, so as to improve the mean average precision of beets and weeds. Furthermore, considering that the traditional convolutional kernel would restrict the geometric modeling ability of the model, deformable convolution was adopted in the feature extraction layer. Moreover, online hard example mining was introduced to excavate the hard negative samples in the detection process to retrain the misidentified samples. Through the aforementioned improvement measures, the average precision of beets and weeds were 84.8% and 93.2% respectively, and the mean average precision was increased from 82.3% to 89.0%. Compared with the original model, the detection accuracy was

improved by approximately 7 percentage points. Our results demonstrate that the performance of the optimized model is not only greatly improved, but the quantity of the model parameters is also maintained at a reasonable amount, provide a theoretical basis for the subsequent development of intelligent weed control robots.

Performance evaluation of the model

Precision and recall are widely used in the field of information retrieval. As with all machine learning problems, in order to calculate precision and recall, it is necessary to explain the following: True Positive (TP), the number of positive classes predicted as positive classes; False Positive (FP), the number of negative classes predicted as positive classes (error rate); True Negative (TN), the number of negative classes predicted as negative classes; and False Negative (FN), the number of positive classes predicted as negative classes (missing rate). Based on this, we define the following:

$$Precision = \mathrm{TP}/(\mathrm{TP}+\mathrm{FP}) \quad (3)$$

$$Recall = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN}) \quad (4)$$

The average precision (AP) is then calculated as follows:

$$\mathrm{AP} = \int_0^1 p(r)d(r) \quad (5)$$

where p represents $Precision$, r represents $Recall$, and p is a function taking r

as a parameter. The mean average precision (mAP) equals the average of all the AP categories, and is determined as follows:

$$\mathrm{mAP} = \frac{\sum \mathrm{AP}}{\mathrm{N}(classes)} \quad (6)$$

where $classes$ represents the detected objects, and N is the number of all categories of objects to be detected.

The impact of multi-modality fusion images on mAP

Since the algorithm proposed in the current study only supports three-channel images as the input data, while the NIR image was composed of a single channel, RGB and multi-modality images (three-channel fusion images) were used in all ablation experiments for verification. As reported in Table 1, the classical region-based fully convolutional network (R-FNC) model and our proposed model exhibited higher detection accuracies on the fusion data set compared with the RGB data set. This is attributed to the high sensitivity of the shallow features of beets and weeds in visible images under the low levels of light, as well as the complex environment. However, the near-infrared image is able to depict the thermal radiation of the target objects. The surface reflectivity of the object was completely different to that of the background, and was thus more robust to variations in light[26]. However, as the near-infrared images are of a low spatial resolution and contain limited texture information, the visible and NIR images were fused into a multi-modality image. These fusion images were sent to the CNN for training. The final

detection accuracy of the model greatly improved with the use of the fusion data set, and the detection system was also more robust.

Table 1. The APs and mAP for different types of images

The detection results were visualized to further demonstrate the performance of multi-modality images. As can be seen from figure 7, image triplet (a) demonstrates the detection result of the improved R-FCN model on the RGB data set, while image triplet (b) depicts the result of the same model using the multi-modality data set. Due to the fusion of the near-infrared and visible images for the multi-modality image, the feature information of beet seedlings and weeds were able to be better characterized under the terrible light and complex field backgrounds. Thus, the detection performance of the multi-modality fusion images was better than that of the improved R-FCN model. Hence, the subsequent ablation experiments were conducted using on the fusion data set.

Figure 1. Testing results using different types of the images

The impact of deformable convolution on mAP

Models 2 and 3 used deformable convolutions, while models 0 and 1 implemented traditional convolutions (Table 2). Compared with the traditional convolution model, the deformable convolution model was able to improve the mAP of beets and weeds by 3-4 percentage points. This can be attributed to the offset variables of the deformable convolution kernel, allowing for the feature expression of the CNN to automatically adapt

193 to changes in the morphological of the target object.

194

195 **Table 2. Model parameter settings and mAP**

196

197 Figure 8 depicts the convolution results for the multi-modality data set. The deformable
198 convolution results (image triplet (b)) exhibited a higher detection accuracy and a lower
199 missing rate than the traditional convolutional model (image triplet (a)). This is a result
200 of the ability of the deformable convolution model to adaptively detect irregular
201 geometric edges of seedlings when detecting smaller target objects.

202

203 **Figure 2. Testing results using deformable convolution and traditional convolution**

204

205

206 *The impact of OHEM on mAP*

207 OHEM retrains the hard samples with large loss values, as such samples may lead to the
208 misclassification of weeds and beets. Compared to other models without OHEM, the
209 detection accuracy of models 1 and 3 with the OHEM algorithm were improved (Table
210 2). This indicates that the OHEM method can suppress the simple samples and the small
211 number of samples, making the training process more efficient. In addition, the OHEM
212 algorithm also eliminated several heuristics and hyper-parameters by automatically
213 selecting hard examples, thus simplifying the training process[27].

214

215 *Detection results of the optimal model*

In order to verify the prediction results of the optimal model under the actual field environment, six images of beets and weeds were selected from the test set. As demonstrated in figure 9, with the use of deformable convolution and multi-modality fusion images, our improved model was able to maintain a high detection accuracy. This indicated that the optimal proposed model exhibits strong generalization and robustness abilities for the detection of beet seedlings and weeds under the poor light and complex field backgrounds.

Figure 3. Detection results of the optimal model

Discussion

In modern agriculture, efficient detection and identification of weeds and crops is an important step to improve agricultural efficiency. As an cutting-edge research method, deep learning is more and more widely used in precision agriculture. In the research of automatic weeding machine, effective image preprocessing is the first step for the machine to complete the recognition of images under different lighting conditions. The high detection accuracy means that it can realize the accurate distinction between weeds and crops, reduce the artificial input and improve the agricultural efficiency.

In this study, we successfully achieved the fusion of RGB images and near-infrared images, which confirmed that the accurate detection of beet plants and weeds can still be achieved in the absence of light conditions. The model was further improved by deformable convolution and ohem.

Conclusion

This study proved that, through the fusion of RGB image and near-infrared image, the improved deep learning model R-FCN can achieve better detection results. Aim to get a better detection result, online hard example mining was introduced to excavate the hard negative samples in the detection process for the retraining of misidentified samples. Deformable convolution can also improve the effect of feature extraction. However, there are still a lot of work to be improved, such as how to further improve the accuracy of the model, whether the model can be compressed to the mobile portable robot. Therefore, it is necessary to carry out further research

Materials and methods

Data source

In order to investigate the detection and identification of beets and weeds in complex backgrounds, images of beets and weeds were collected at the University of Bonn, Germany, in 2016. For more details about the dataset, see the link and study below (http://www.ipb.unibonn.d-e/data/sugarbeets2016/). The images were collected via a multi-modality camera (JAI AD-130GE), equipped with two high-sensitivity CCD multispectral sensors of 1.3 million pixels. The camera can simultaneously collect visible (400nm~650nm) and near infrared (NIR) (760nm~1000nm), with an output image size of 1296 × 1296 pixels[17]. The dataset contains a total of 2,093 images of beets and weeds at different growth stages (Figure.1). In the process of data acquisition, beet seedlings and weeds with different levels of maturity under varying angle transformations were consider

for the image acquisition. Moreover, the same plant (beet and weed) was imaged multiple times under different ranges of overlap and occlusion between beets and weeds.

Figure 4. Examples of seedling dataset for weeds and beets

Multi-modality image fusion

In general, object detection methods based on deep learning aim to understand the distribution of basic data via a large amount of training data and subsequently induce the optimizer to adjust the parameters of the network[18]. At present, RGB images on weeds and beets are commonly used to train deep learning models. However, RGB images are sensitive to variations in light, resulting in the loss of important information on the shape, color and texture of target objects[19]. Therefore, the performance of such models is poor under complex backgrounds at night.

In order to solve this problem, in the current study, the NIR and visible images of beets and weeds were fused into multi-channel images. In particular, multi-modality image fusion spatially registers the data of the same image from different sources, and subsequently combines the information in each image to generate an integrated data set of all the images[20].

Deep fusion algorithm frame diagram

Figure 5. Deep fusion method for multi-modality images

Essentially, the deep fusion algorithm extracts features from a denseblock composed of convolutional layers, and then sends the visible and near-infrared images to a fusion layer for pixel-level superposition. These fusion images are reconstructed through a decoding network also composed of convolutional layers to obtain the three-channel multi-modality images. As shown in figure 2, the encoding network consists of two sections: C1 and denseblock. In the C1 section, a 3×3 convolutional kernel is used to extract the rough features of the image. The denseblock section then uses three convolutional layers (the output of each layer is cascaded as the input of the next layer) to extract the high-level abstract semantic features of the image. The denseblock adopted in this research can retain image features as much as possible to ensure that all significant features can be used in the fusion strategy. In the fusion layer, $l_1 - norm$ fusion strategy is selected to fuse the visible and near-infrared images. Finally, four convolutional layers (3×3 convolutional kernel) are used to reconstruct the final fusion image in the decoding network. What's more, in this framework, the size of the input and output images is 1296 × 1296 pixels, and the number of feature mapping channels per convolutional layer is 16. More details of the deep fusion algorithm can be found in the study[21].

Multi-modality fusion image

Figure 3 presents the three-channel multi-modality image obtained via the depth fusion algorithm. Following the fusion of the data, the LabelImg software (https://tzutalin.github.io/Lab-elImg/) was used by experts in the agricultural field to label

the beet seedlings and weeds based on the PASCAL Visual Object Classes Challenge[22]. A python script was then used to randomly divide the images and the corresponding tag files into training and testing sets at a ratio of 4:1.

Figure 6. Examples of RGB and multi-modality images. The image triplet (a) shows the original visible dataset, and the image triplet (b) shows the corresponding fusion dataset.

Improved weeds and beets detection model

Deformable Convolution

Recently, the use of CNNs has made significant breakthroughs in many vision applications. However, due to the regular grid sampling and the fixed geometric structure in traditional convolution methods, it is difficult for networks to deal with geometric deformations. The adaptability of the existing models to process the geometric deformation of objects almost comes from the diversity of the data itself, and there is no mechanism to adapt to geometric deformation in the model. Thus, the ability of geometric transformation modeling is limited, and it cannot be adjusted adaptively according to the image content[23]. In order to overcome this limitation, a new module, deformable convolution, was adopted to improve the modeling ability of CNN for transformations in the current study. More specifically, an offset variable is added to the position of each sampling point in the convolutional kernels. The kernel with the offset variable can then be sampled randomly near the current position, and is thus not restricted to the previous regular grid points. Moreover, the offset variable can be obtained by learning within the

target task without the need of an additional monitoring signal, improving the traditional convolution.

Figure 7. Schematic diagram of deformable convolution. This figure shows the sampling methods of traditional convolution and deformable convolution with convolution kernel size 3×3. Figure 4(a) demonstrates the regular sampling grid (green points) of traditional convolution, while (b) present the deformed sampling locations (black points) with augmented offsets (blue arrows) in deformable convolution. Then (c) and (d) are special cases of (b), showing that deformable convolution generalizes scale, aspect ratio and rotation transformations.

Figure 8. Illustration of 3×3 deformable convolution

Figure 5 presents the internal structure of deformable convolution. First, the displacement required for deformable convolution is obtained through the output of a small convolutional layer, and the displacement is then applied to the convolutional kernel in order to achieve the effect of deformable convolution. This operation is able to add the offsets to the regular grid sampling locations in the standard convolution, thus enabling the free form deformation of the sampling grid. The offsets are learned from the preceding feature maps via additional convolutional layers.

Traditional convolution consists of two steps: 1) Sampling using a regular grid R over the input feature map x; and 2) the summation of sampled values weighted by w. For each location p_0 on the output feature map y, traditional convolution is then performed as follows：

$$y(p_0) = \sum_{p_n \in R} w(p_n) \cdot x(p_0 + p_n) \quad (1)$$

where the grid R defines the receptive field size and dilation of a 3×3 kernel with

dilation 1, $R = \{(-1,-1),(-1,0),\ldots,(0,1),(1,1)\}$, and p_n enumerates the locations in R.

In deformable convolution, the grid R has the offsets $\{\Delta p_n | n = 1,2,\ldots,N\}$, and $N = |R|$, thus, formula (1) becomes:

$$y(p_0) = \sum_{p_n \in R} w(p_n) \cdot x(p_0 + p_n + \Delta p_n) \quad (2)$$

As the offset Δp_n is typically fractional, formula (2) is implemented via bilinear interpolation.

The network structure of the improved model

Figure 9. Key idea of the improved R-FCN for weeds and beets detection.

The network structure of the improved model is described as follows. Following the input of the image, feature extraction was performed, resulting in a feature map of $k^2 \times (C+1)$ dimensions. The region proposal network (RPN)[24] was then used to extract the regions of interest (ROIs) in the feature map, with C denoting the number of categories. The extracted ROIs were divided into $k \times k$ regions, with k generally equal to 3, corresponding to 9 regions: top-left, top-center, and bottom-right. Finally, the score of each region was determined by a pooling operation, and the output feature vector of the ROIs were then obtained by voting. This vector was subsequently used for classification and regression of the weeds and beets.

Experimental environment

Equipment and platform

The Ubuntu 16.04 system is used as the operating platform, and MXNet is adopted as the deep learning framework to train the network. The computer memory of the system is 32GB, with a 3.6 GHZ i7-9700k CPU processor. Additionally, the 11 GB GeForce GTX1080Ti GPU with Pascal architecture was used.

Model parameter setting

In order to reduce variation of parameter updates and to stabilize the convergence of the model, a mini-batch stochastic gradient descent (SGD) was used to train the network[25]. The parameters were set as follows: the number of each mini-batch of samples was 128, the momentum factor was fixed to 0.9, and the weight attenuation factor was 0.0005 to avoid over-fitting. Finally, the gradient descending learning rate was applied to all layers of the network, and was gradually reduced in stages to 0.1 times of the current learning rate. Additionally, the initial learning rate during the training process was set to 0.005, and the model was iterated for 100 epochs.

Acknowledgements

We are very grateful to Professor Sun Jun, Wu Xiaohong and Shen Jifeng for their suggestions to the manuscript.

Authors' contributions

Yang Yu and He Xiaofei conceived and designed research. Yang Yu conducted experiments. All authors read and approved the final manuscript.

Funding

This work is partially supported by a project funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD-2018-87), Synergistic Innovation Center of Jiangsu Modern Agricultural Equipment and Technology (4091600002). Project of Faculty of Agricultural Equipment of Jiangsu University (4121680001)

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding authors on reasonable request.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. Bauer M V, Marx, C, Bauer F V, Flury D M, Ripken T, Streit B. Thermal weed control

technologies for conservation agriculture—a review. Weed Res.2020;00:1–10. https://doi.org/ 10.1111/wre.12418

2. Hu K, Coleman G, Zeng S, Wang Z, Walsh M. Graph weeds net: a graph-based deep learning method for weed recognition. Comput Electron Agric 2020;174, 105520. https://doi.org/ 10.1016/j.compag.2020.105520

3. Saichao Gao, Guobin Wang, Yangyang et al. Water-soluble food dye of allura red as a tracer to determine the spray deposition of pesticide on target crops. Pest Manag Sci. 2019;75:2592-2597. https://doi.org/ 10.1111/wre.12418

4. Gai J, Tang L, Steward B L. Automated crop plant detection based on the fusion of color and depth images for robotic weed control. J. Field Robot. 2019; 37:35-52. https://doi.org/ 10.1002/rob.21897

5. Raja R, Nguyen TT, Slaughter DC. Fennimore SA. Real-time robotic weed knife control system for tomato and lettuce based on geometric appearance of plant labels. Biosyst Eng.2020;194:152-164. https://doi.org/ 10.1016/j.biosystemseng.2020.03.022

6. Sun, J., He, X., Tan, W., Wu, X., Shen, J., Lu, H. Recognition of crop seedling and weed recognition based on dilated convolution and global pooling in CNN. Trans Chin. Soc Agric Eng. 2018;34(11):159-165. https://doi.org/ 10.11975/j.issn.1002-6819.2018.1-1.020

7. Zhao P, Wei Z. Weed Recognition in Agricultural Field Using Multiple Feature Fusions. Trans Chin Soc Agric Mach. 2014;45(3):275-281. https://doi.org/10.60-41/j.issn.1000-1298.2014.03.045

8. Yan B. Identification of Weeds in Maize Seedling Stage by Machine Vision Technology. J. Agric. Mechanization Res. 2018;40(3): 212-216. https://doi.org/10.3969/j.issn.1003-188X.2018.03.043

9. Bakhshipour A, Jafari A, Nassiri S M, Zare D. Weed segmentation using texture features extracted from wavelet sub-images. Biosyst Eng. 2017; 157: 1-12. https://doi.org/10.1016/j.biosyste-mseng.2017.02.002

10. Nogueira K , Penatti O A B, Santos J A D. Towards Better Exploiting Convolutional Neural Networks for Remote Sensing Scene Classification. Pattern Recognition. Pattern Recognit. 2017;61:539-556. https://doi.org/10.1016/j.patcog.2016.07.001

11. Palanisamy T, Sadayan G, Pathinetampadiyan N. Neural network-based leaf classification using machine learning. Concurrency Computat Pract Exper. 2019;(15):e5366.

12. Rehman T U, Zaman Q U, Chang Y K, Schumann A W, Corscadden K W. Development and field evaluation of a machine vision based in-season weed detection system for wild blueberry. Comput Electron Agric. 2019; 162:1-13. https://doi.org/ 10.1016/j.compag.2019.03.023

13. Sun, J., Tan, W., Wu, X., Shen, J., Lu, B., Dai, C. Real-time recognition of sugar beet and weeds in complex backgrounds using multi-channel depth-wise separable convolution model. Trans Chin Soc Agric Eng. 2019; 35(2). https://doi.org/10.11975/j.issn.1002-6819.2019.12.000

14. Jiang H, Wang P, Zhang Z, Mao W, Zhao B, Qi P. Fast Identification of Field Weeds

458 Based on Deep Convolutional Network and Binary Hash Code. Trans Chin Soc Agric
459 Eng. 2018;49(11):37-45. https://doi.org/ 10.6041/j.issn.1000-1298.2018.11.004
460 15. Andrea, C. C., Daniel, B. B. M., Misael, J. B. J. Precise weed and maize classification
461 through convolutional neuronal networks. IEEE Second Ecuador Technical Chapters
462 Meeting (ETCM). 2017; pp. 1-6. https://doi.org/10.1109/ETCM.2017.8247469
463 16. Joseph E Hunter, Travis W Gannon, Robert J Richardson, Fred H Yelverton, Ramon
464 G Leon. Integration of remote-weed mapping and an autonomous spraying unmanned
465 aerial vehicle for site-specific weed management. Pest Manag Sci. 2020; 76: 1386–
466 1392. https://doi.org/ 10.1002/ps.5651
467 17. Milioto A, Lottes P, Stachniss C. Real-time semantic segmentation of crop and weed
468 for precision agriculture robots leveraging background knowledge in CNNs. Proc Intl
469 Conf on Robotics and Automation. 2018; 2229-2235. https://doi.org/1-
470 0.1109/ICRA.2018.8460962
471 18. Shin H C, Roth H R, Gao M, Lu L, Xu Z, Nogues I. Deep convolutional neural
472 networks for computer-aided detection: CNN architectures, dataset characteristics and
473 transfer learning. IEEE transactions on medical imaging. 2016;35(5):1285-1298.
474 https://doi.org/10.1109/TM-I.2016.2528162
475 19. Ying, Z., Li, G., Ren, Y., Wang, R., Wang, W. A new image contrast enhancement
476 algorithm using exposure fusion framework. Int. Conf on Comput Analysis of Images
477 and Pattern. 2017; 36-46.
478 20. Ren X, Meng F, Hu T, Liu Z, Wang C. Infrared-Visible Image Fusion Based on

Convolutional Neural Networks. Int Conf on Intell Sci and Big Data Eng. 2018; 301-307. https://doi.org/10.1007/978-3-030-02698-1_26

21. Li H, Wu X. DenseFuse: A Fusion Approach to Infrared and Visible Images. IEEE Trans on Image Processing. 2019; 28(5):2614-2623. https://doi.org/10.1109/TIP.2018.2887342

22. Everingham M, Winn J. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Development Kit. Int J of Comput Visi. 2006;111(1):98-136. https://doi.org/10.1-007/s11263-014-0733-5

23. Dai J, Qi H, Xiong Y, Li Y, Zhang G, Hu H. Deformable Convolutional Networks. Proc IEEE int conf on comput visi. 2017; 764-773. https://doi.org/10.1109/ICCV.2017.89

24. Ren, S., He K., Girshick, R., Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems. 2015;91-99.

25. Zhou Y, Xu T, Deng H, Miao T. Real-time recognition of main organs in tomato based on channel wise group convolutional network. Trans Chin Soc Agric Eng. 2018;34(10):153-162. https://doi.org/10.11975/j.issn.1002-6819.2018.10.019

26. Raghavendra, R., Dorizzi, B., Rao, A., Kumar, G.H.. Particle swarm optimization based fusion of near infrared and visible images for improved face verification. Pattern Recognit 2011;44(2). https://doi.org/10.1016/j.patcog.2010.08.006

27. Shrivastava A, Gupta A, Girshick R. Training region-based object detectors with

500 online hard example mining. Proc IEEE Conf Comput Vis and Pattern Recognit.
501 2016;761-769. http:// dx.doi.org/10.1109/CVPR.2016.89
502

Figures

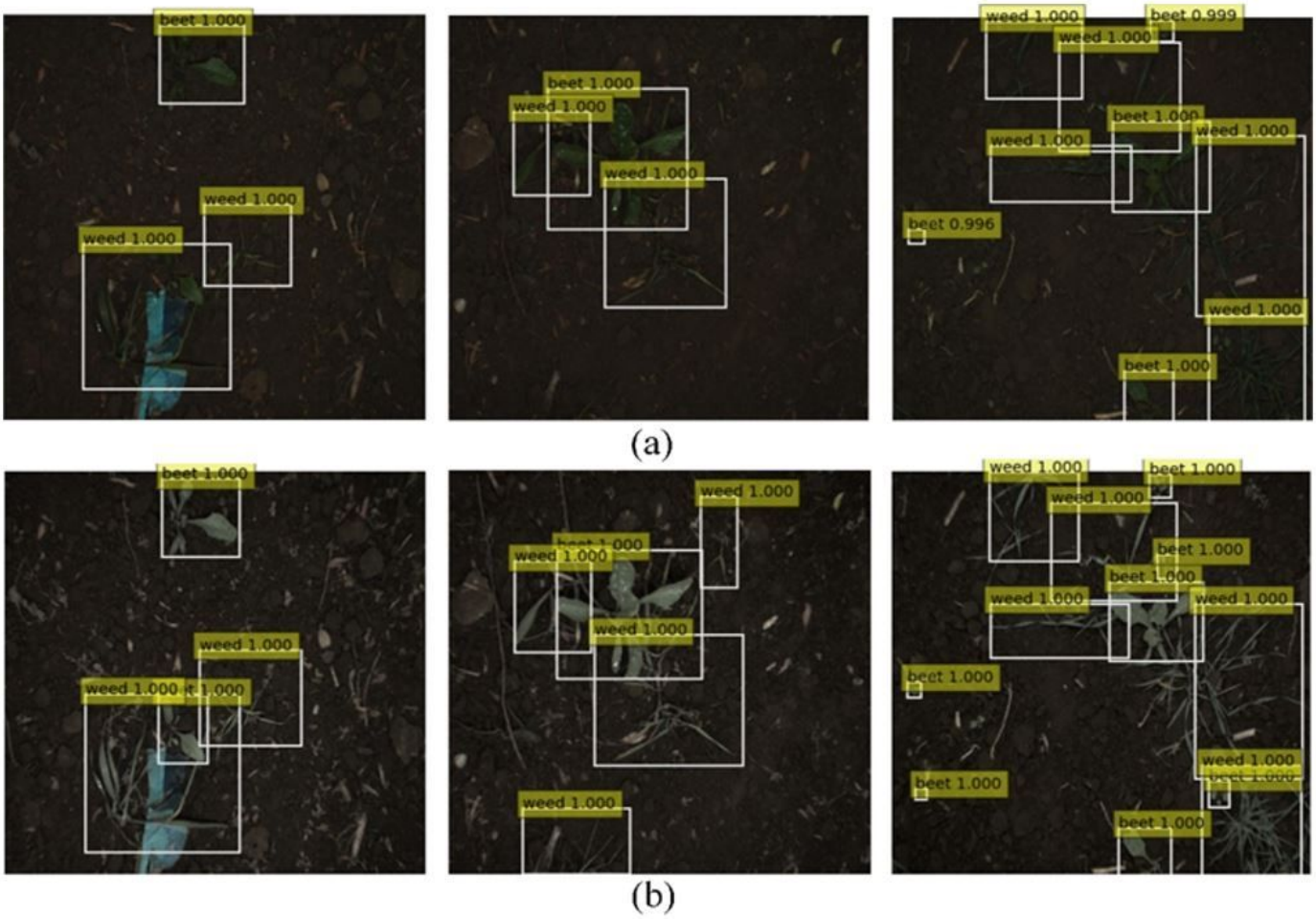

Figure 1

Testing results using different types of the images

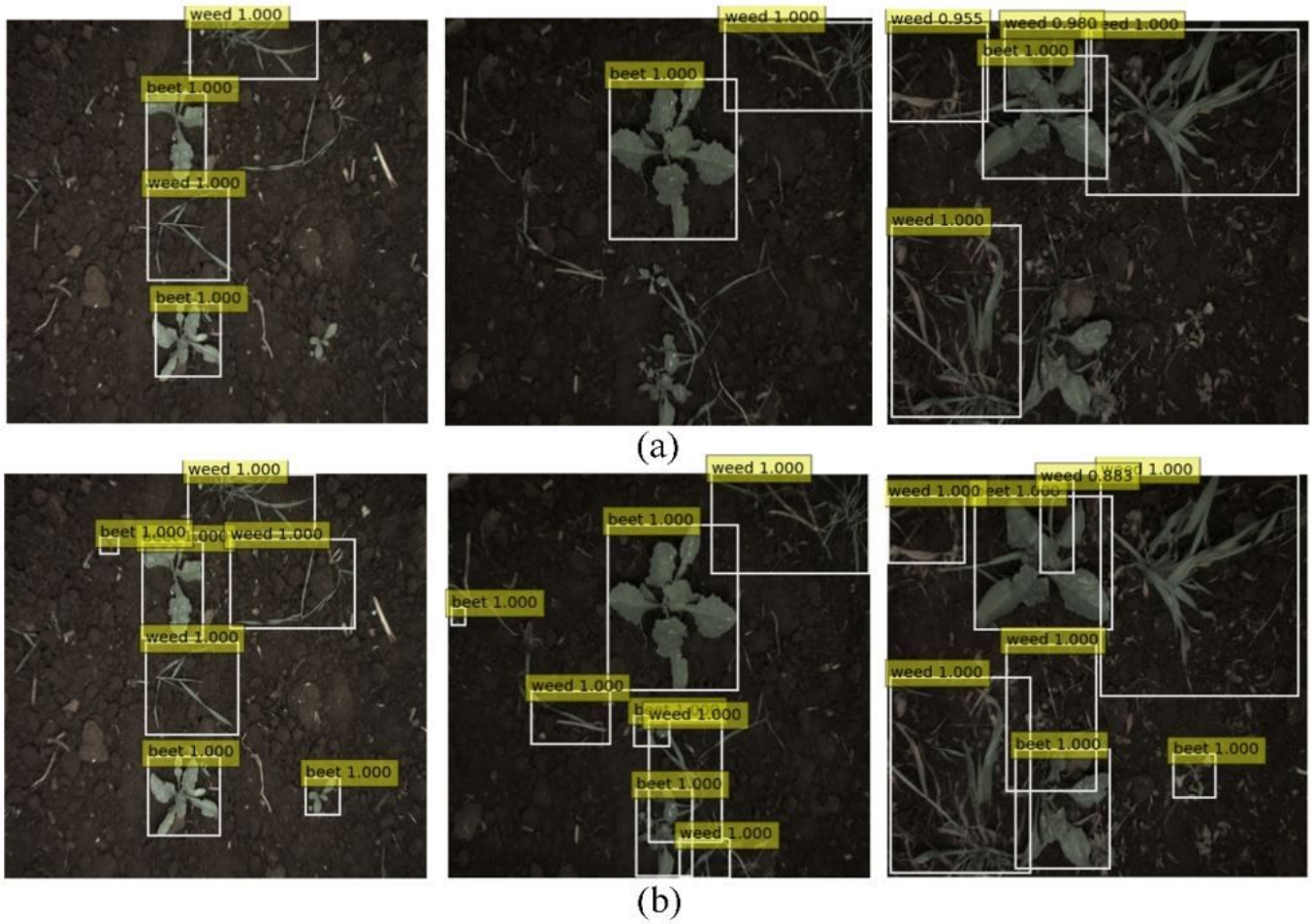

Figure 2

Testing results using deformable convolution and traditional convolution

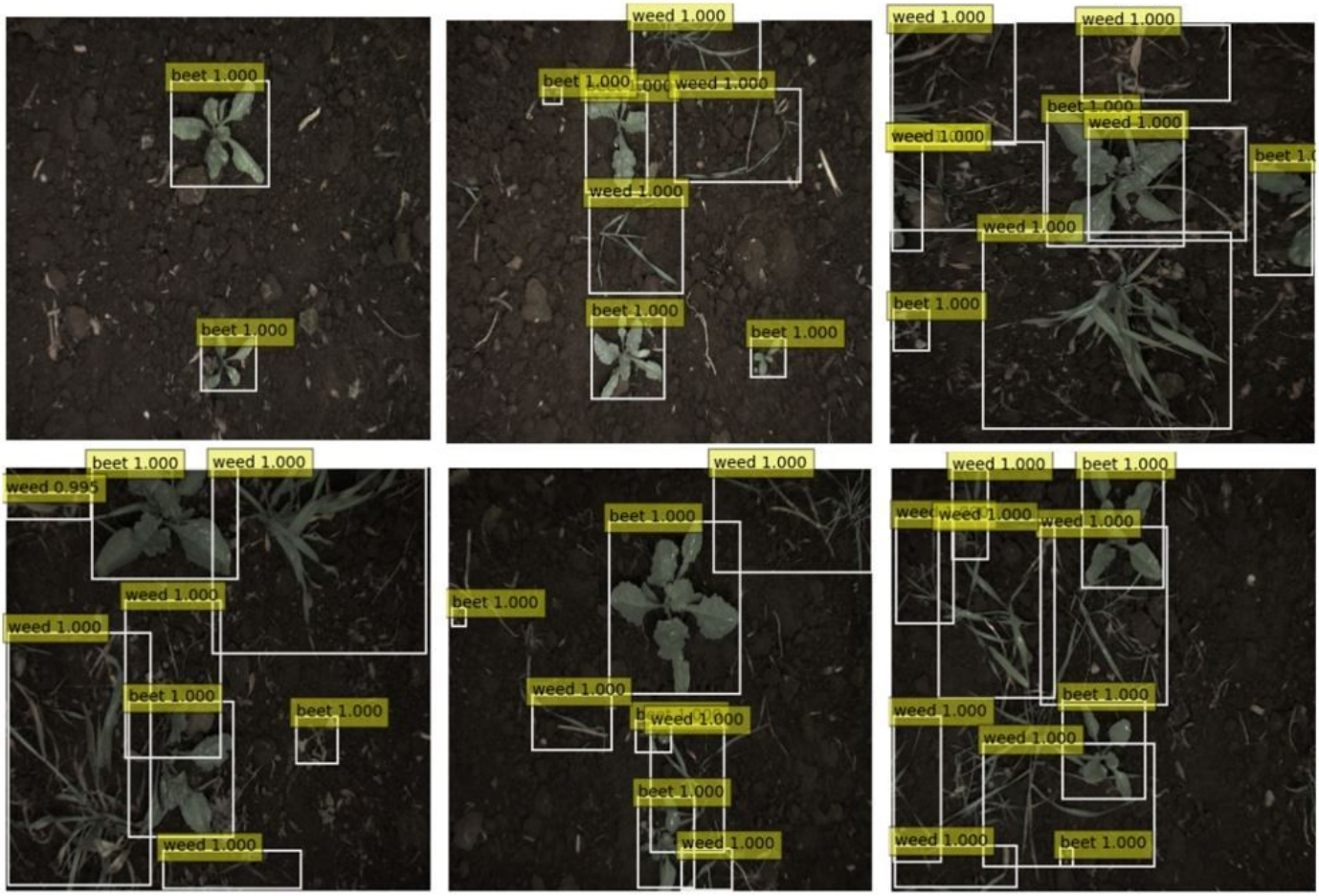

Figure 3

Detection results of the optimal model

Figure 4

Examples of seedling dataset for weeds and beets

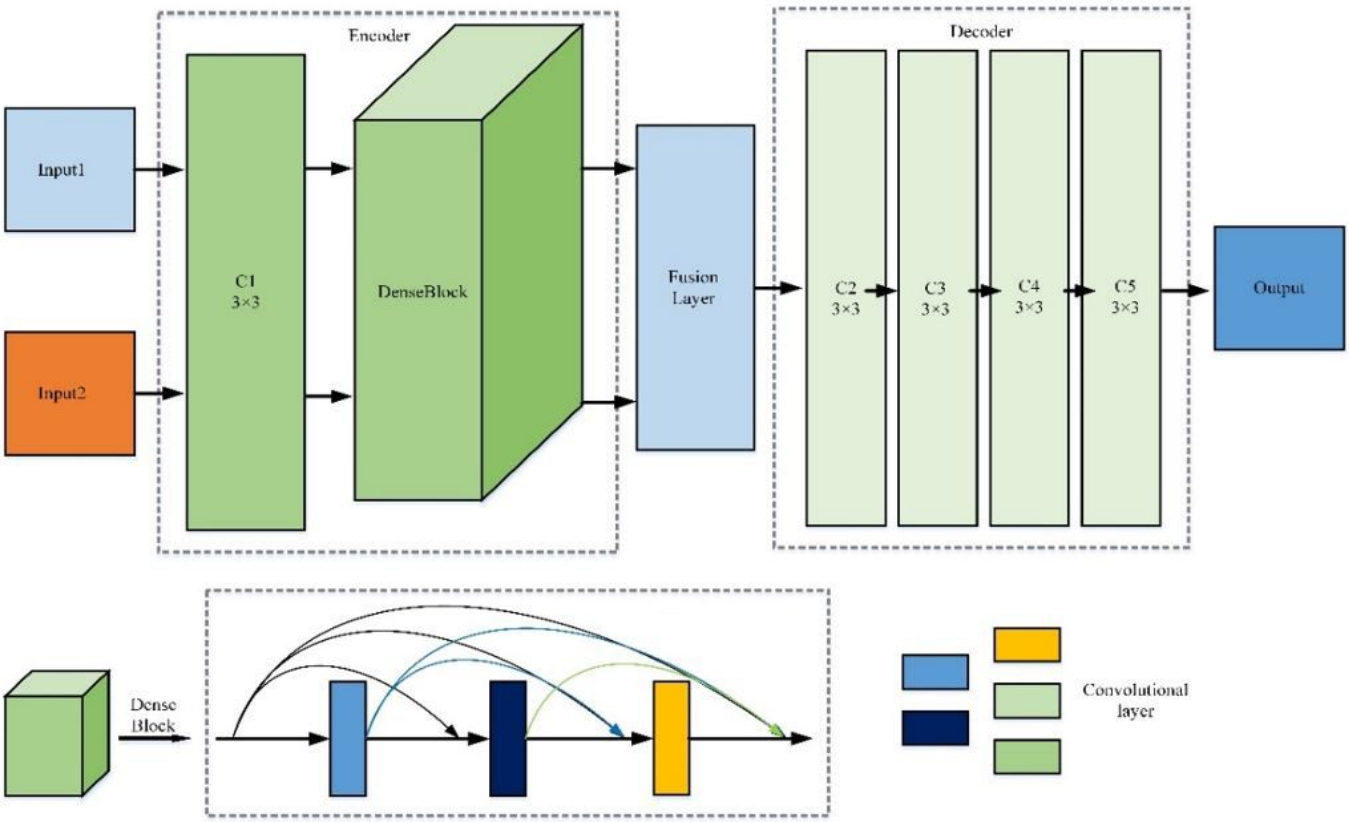

Figure 5

Deep fusion method for multi-modality images

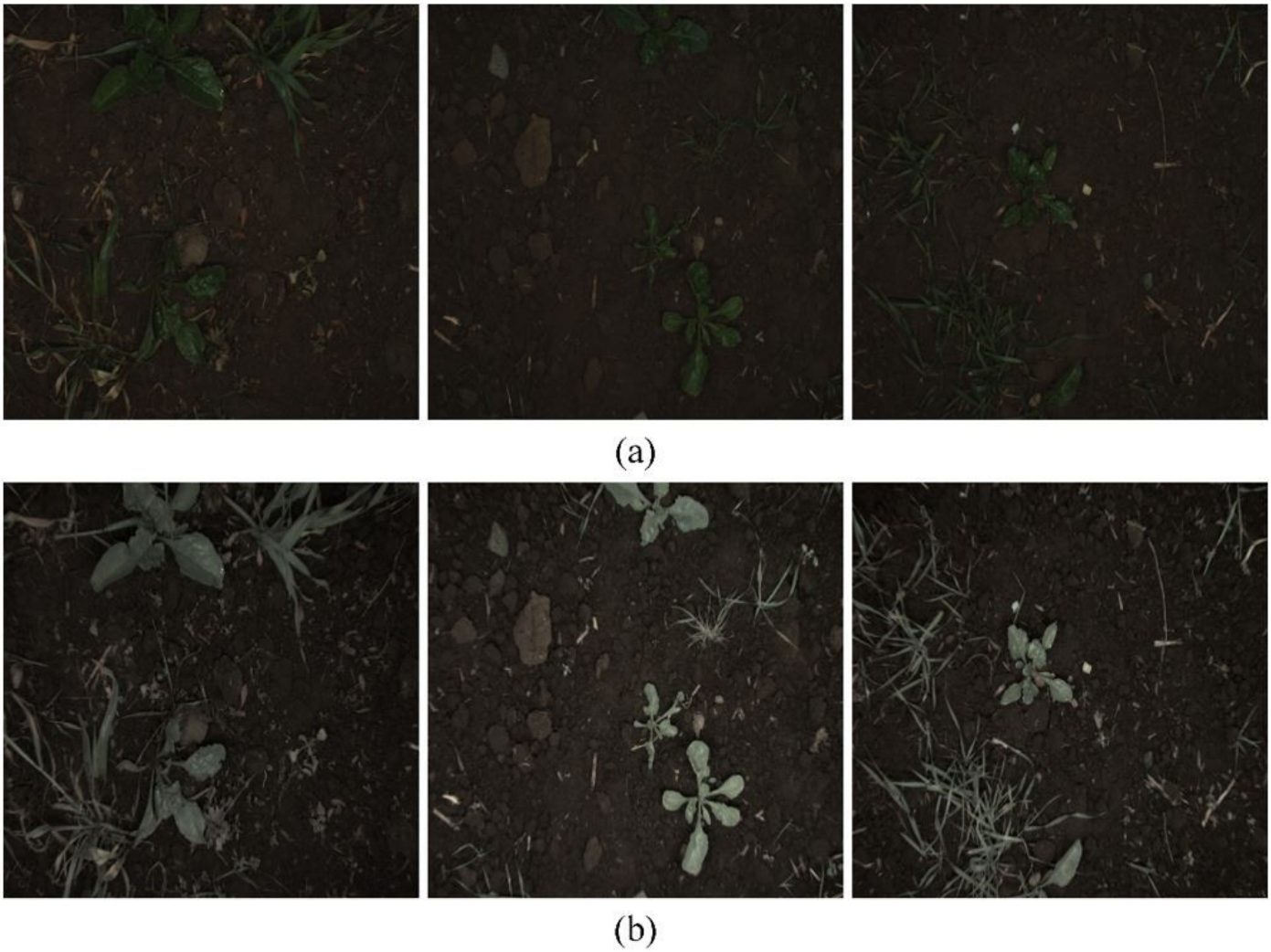

Figure 6

Examples of RGB and multi-modality images. The image triplet (a) shows the original visible dataset, and the image triplet (b) shows the corresponding fusion dataset.

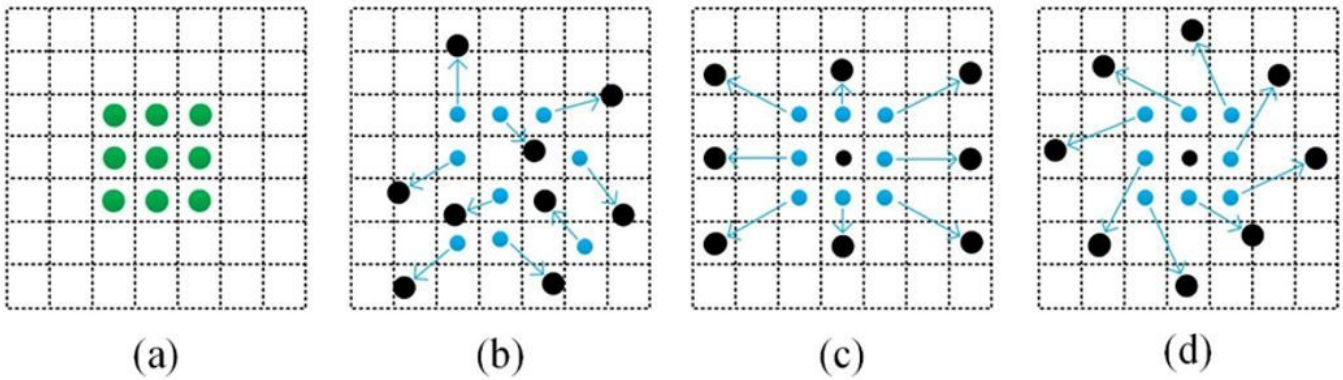

Figure 7

Schematic diagram of deformable convolution. This figure shows the sampling methods of traditional convolution and deformable convolution with convolution kernel size 3×3. Figure 4(a) demonstrates the regular sampling grid (green points) of traditional convolution, while (b) present the deformed sampling locations (black points) with augmented offsets (blue arrows) in deformable convolution. Then (c) and (d) are special cases of (b), showing that deformable convolution generalizes scale, aspect ratio and rotation transformations.

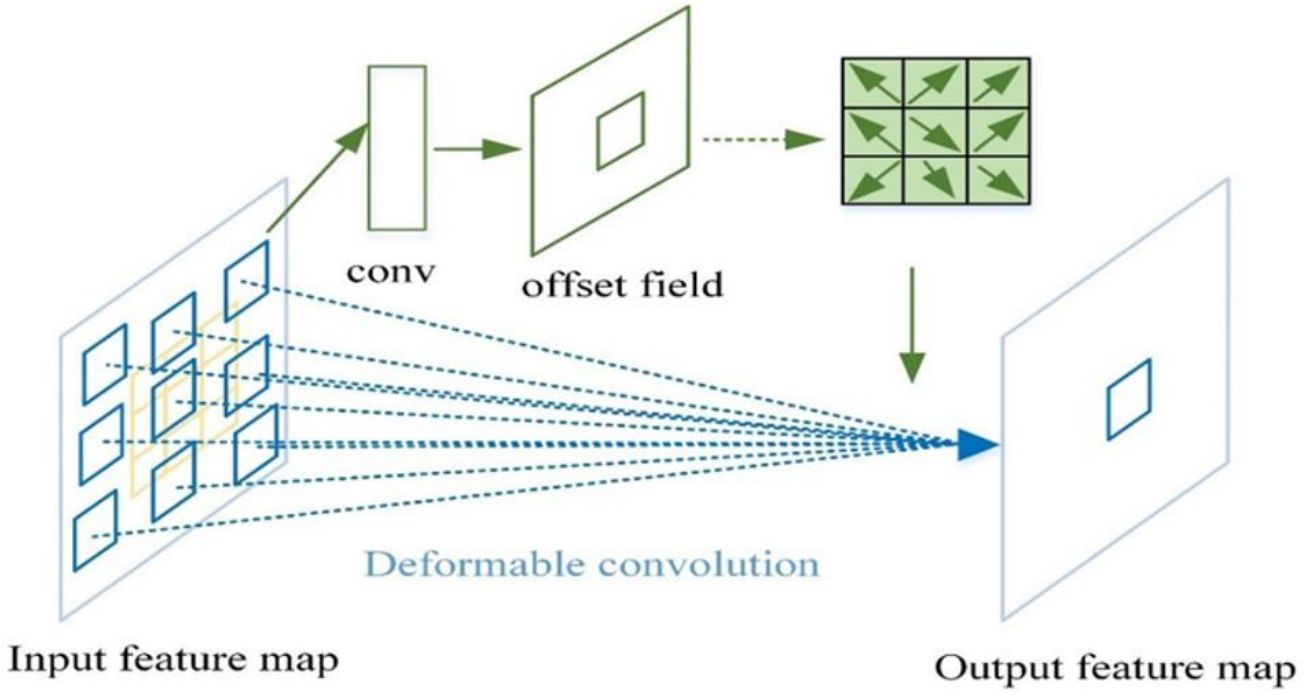

Figure 8

Illustration of 3×3 deformable convolution

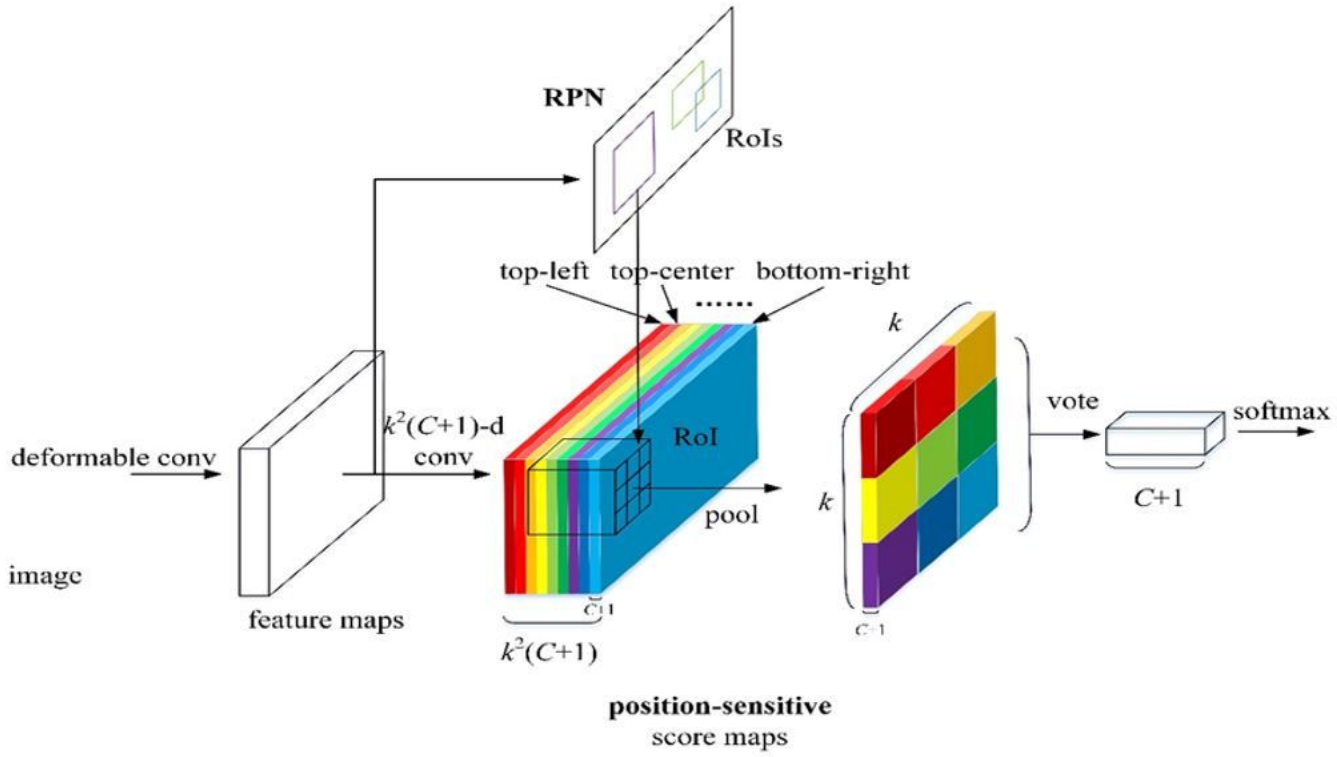

Figure 9

Key idea of the improved R-FCN for weeds and beets detection.