Computers and Electronics in Agriculture 190 (2021) 106418

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Original papers

Data augmentation for deep learning based semantic segmentation and crop-weed classification in agricultural robotics

Daobilige Su [a], He Kong [b,*], Yongliang Qiao [b], Salah Sukkarieh [b]

[a] College of Engineering, China Agricultural University, Beijing, China
[b] Australian Centre for Field Robotics, The Rose Street Building J04, The University of Sydney, NSW 2006, Australia

ARTICLE INFO

Keywords:
Data augmentation
Deep learning
Semantic segmentation
Agricultural robot
Crop weed classification

ABSTRACT

Deep learning methods such as convolutional neural networks (CNN) have become popular for addressing crops and weeds classification problems in agricultural robotics. However, to have satisfactory performance and avoid overfitting, training deep neural nets typically requires thousands of labeled images. This leads to tedious pixelwise labeling for semantic segmentation. In this paper, we hinge on the recent development in data augmentation and utilize the concept further for semantic segmentation and classification of crops and weeds. To be specific, we propose a novel data augmentation framework, based on the random image cropping and patching (RICAP) method, which is originally designed to augment data for generic image classification. The proposed framework introduces novel enhancements to the original RICAP so that it can be effectively used for data augmentation of semantic segmentation tasks. We evaluate the proposed methodology on two datasets from different farms. Comprehensive experimental evaluations and ablation studies show that the proposed framework can effectively improve segmentation accuracies, and the enhancements made over the original RICAP actually contribute to the performance gain. On average, the proposed method increases the mean accuracy and mean intersection over union (IOU) of the deep neural net with the conventional data augmentation (random flipping, rotation and colour jitter) from 91.01 to 94.02 and from 63.59 to 70.77 respectively for Narrabri dataset, and from 97.99 to 98.51 and from 74.26 to 77.09 respectively for Bonn dataset. The limitation of the proposed method, especially when a large number of training data is available, has also been discussed.

1. Introduction

Robotic applications in precision agriculture have become a popular topic recently. Deploying robots in agricultural applications has the potential to significantly reduce the labor cost of repetitive activities that are traditionally carried out by humans. The agricultural tasks that researchers have been trying to automate include, but are not limited to, weeding (McCool et al., 2018; Su et al., 2021), plant detection and yield estimation (Chlingaryan et al., 2018), fertiliser or pesticide application (Mortensen et al., 2018).

Specifically, in the field of precision plant level weeding and fertilising, the robot needs to classify crops and weeds on images acquired from its onboard camera. This forms a semantic segmentation problem, and various methods based on deep CNNs have been proposed recently to accomplish such a task and showed remarkable performance (Milioto and Stachniss, 2018; Potena et al., 2016). Semantic segmentation classifies each pixel in the image to be one of the existing classes. Therefore it provides more precise locational information of the crop and weed. To render satisfactory performance and avoid overfitting, these deep neural net based methods, with a large number of tuning parameters, typically require a few thousands to tens of thousands of images to train.

However, producing ground truth pixelwise semantic labels for images acquired from agricultural fields is significantly more challenging and time consuming than images of household or urban scenes (Bosilj et al., 2019). This is due to the fact that crops and weeds on the images typically have irregular shapes, and simple polygon based labeling tools are not exceptionally helpful in this situation. This results in manual labeling of images normally taking 30–60 min per image, which makes generation of tens of thousands of images with ground truth labels very time consuming or infeasible (Sa et al., 2018). This is also the main reason why publicly available datasets for crop weed segmentation

* Corresponding author.
E-mail addresses: sudao@cau.edu.cn (D. Su), he.kong@sydney.edu.au (H. Kong), yongliang.qiao@sydney.edu.au (Y. Qiao), salah.sukkarieh@sydney.edu.au (S. Sukkarieh).

https://doi.org/10.1016/j.compag.2021.106418
Received 4 August 2020; Received in revised form 11 August 2021; Accepted 21 August 2021
Available online 21 September 2021

(Chebrolu et al., 2017; Sa et al., 2018; Minervini et al., 2016) only contain around one hundred to few hundreds labeled images [1]. Using these data alone, neural nets with deep CNN architecture tend to overfit.

In this paper, we propose a data augmentation method, based on the Random Image Cropping And Patching (RICAP) (Takahashi et al., 2018), so that it can be effectively used for data augmentation of semantic segmentation tasks of agricultural datasets. Moreover, we introduce two novel enhancements to the original RICAP method, which improve the segmentation accuracies of the trained deep neural net further.

The remainder of the paper is organised as follows. A literature review of work related to data augmentation for deep neural nets is provided in Section 2. The proposed method of data augmentation for crop weed semantic segmentation is described in Section 3. Experimental evaluations of the proposed method using two data collected from different farms and discussion of the performance of the method are presented in Section 4. Finally, Section 5 presents the conclusions.

2. Related work

Data augmentation can increase the variety of training samples and prevent overfitting. Similar concepts have been leveraged in the fields of deep learning in recent years, and there are many existing methods in the literature that prove to improve the performance of various neural nets.

In AlexNet (Krizhevsky et al., 2012), the authors use random horizontal flipping and cropping to increase the number of images. Such random horizontal flipping and cropping prevent the deep neural net from overfitting to some specific features and increase variations in images with different orientations. The authors also use principal component analysis to change intensity values of all image channels to increase the variations of images further. A similar colour translation technique is introduced in the reimplementation of ResNet by Facebook AI Research (Facebook, 2019). In the latter work, the authors introduce colour jitters to training images to enlarge the dataset.

With recent deep neural nets going increasingly advanced and the number of parameters of the model growing ever larger (He et al., 2016; Zagoruyko and Komodakis, 2016), new data augmentation methods start to emerge. In (Hinton et al., 2012), data augmentation is carried out by introducing dropout, which brings in disturbances to the original images by dropping certain pixels. The method increases the robustness of the trained model to noisy images. However, such a method contributes mostly to generalisation rather than enriching the training dataset. DeVries and Taylor (2017) proposes a cutout-based data augmentation method, which masks out a random square region in images in every training iteration. Such a method not only makes the trained model robust to noisy images, but also forces the model to learn different parts of objects rather than just the main part of objects. Zhang et al. (2017) has introduced a data augmentation technique that alpha-blends two images into one new training image. The mixup treatment increases the number of training images and behaves like perturbation (Goodfellow et al., 2014), thereby improving robustness of the model to adversarial noises. In (Cubuk et al., 2018), the authors propose AutoAugment, which uses a reinforcement learning framework to find the optimum combination of existing data augmentation techniques. Its remarkable performance in CIFAR-10 benchmark dataset shows the benefits of data augmentation techniques. In (Takahashi et al., 2018), the authors propose RICAP, which randomly crops four images and patches them to create a new training image. When being used for image classification tasks, the RICAP mixes the class labels of the four images, and weights each of them by the area of cropped images. This results in an advantage similar to label smoothing. Zoph et al. (2020) investigated learned data augmentation policy on object detection performance. Their method achieves a larger improvement when the training set is small, or tackling harder tasks of detecting smaller objects and detecting with more precision. Zhong et al. (2020) proposed random erasing, which randomly selects a rectangle region in an image and erases its pixels with random values. It yields consistent improvement over strong baselines in image classification, object detection and person re-identification. Wang et al. (2021) proposed an implicit semantic data augmentation (ISDA) algorithm, that consistently improves the generalization performance of popular deep models. ISDA can be formulated as a novel robust loss function, which is compatible with any deep network using the softmax cross-entropy loss.

Although the methods mentioned above demonstrate performance improvement in various deep neural nets, most of them are not directly applicable to off-the-shelf semantic segmentation deep neural nets. Take the cutout method of (DeVries and Taylor, 2017) as an example, the region being masked out does not have corresponding label since it does not correspond to any of the predefined categories. Similarly, the original RICAP in (Takahashi et al., 2018) is designed for image classification, and is not directly applicable for the semantic segmentation problem.

More specific to data augmentation of agricultural datasets, Di Cicco et al. (2017) propose a data augmentation method for semantic segmentation of crop and weed by synthetically generating training data, which can be used to train a deep segmentation neural net. However, the method requires to know information about the distribution of leaves in all crop and weed species, and high definition of textures on leaves and illumination (Bosilj et al., 2019; Bah et al., 2018). Hall et al. (2017) present an unsupervised method, which clusters the training regions to minimise the number of regions that the domain expert needs to process, so that the effort required in the data annotation process can be reduced. However, their method does not enrich the variety of the training data. Bah et al. (2018) propose an unsupervised labeling method, which exploits line structures of crops on aerial images. However, the method can not tackle images captured by ground field robots, *e.g.* as shown in Fig. 3, in which crops do not appear in lines as they are in aerial images. Liu et al. (2020) proposed the Leaf GAN, which is based on generative adversarial networks to generate images of four different grape leaf diseases for training identification models. Their method works well for image classification task, but does not deal with image segmentation task. A similar approach was also proposed by Cap et al. (2020), which also uses generative adversarial networks with own attention mechanism for leaf disease diagnosis application.

3. The data augmentation method

In this paper, we borrow concepts from the RICAP technique (Takahashi et al., 2018), and propose a novel data augmentation method for semantic segmentation. More specifically, we improve the original RICAP by introducing two novel enhancements to improve segmentation accuracies of the trained neural net.

Firstly, we relax the fixed number of 2×2 patches in the original RICAP to make the number of horizontal and vertical patches more suitable for each individual dataset. Especially, our experiment results (see in Section 4.2 for details) show that, to have satisfactory performance, the ratio of horizontal and vertical patches should be similar to the ratio of horizontal and vertical size of informative part of images, but the total number of patches should not be too large to make the size of each patch much smaller than the object to be segmented. Secondly, when patching images, our method proposes several candidates for each image patch, and select the one which has similar edges along the boundaries of cropped images and ground truth segmentation masks, as detailed in Section 3.2.

[1] In Chebrolu et al. (2017), only 283 labeled images are publicly available originally, which is extended to include around $10K$ labeled images later by exploiting additional NIR channel of images. Moreover, in Sa et al. (2018), 155 labeled images are publicly released. In (Minervini et al., 2016), on average, around one hundred labeled images are available for various tasks.

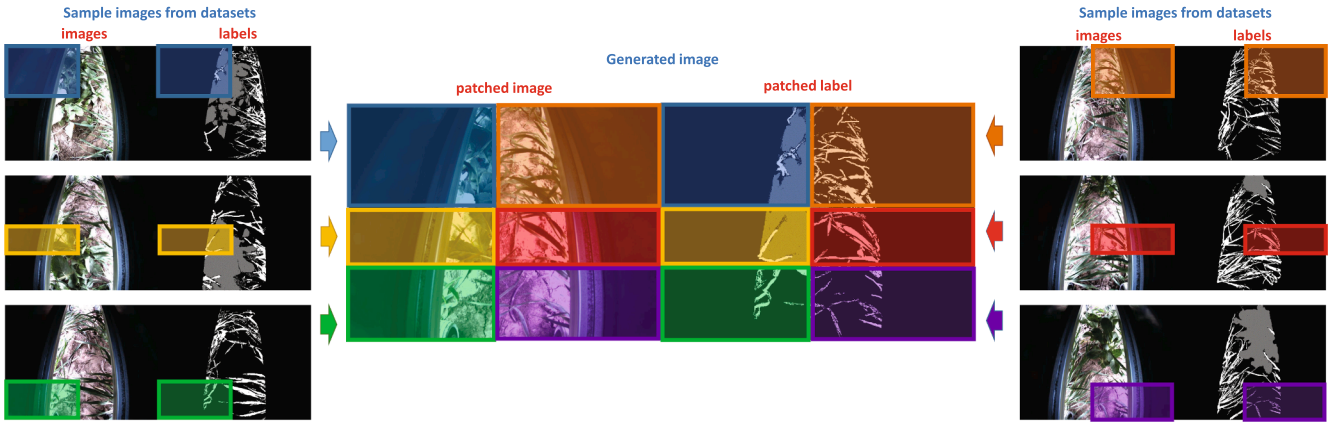


Fig. 1. Illustration of the proposed method. In this example, the image region is divided into 2 parts horizontally and 3 parts vertically. The size of horizontal and vertical slices are random for each generated image. Then 6 images and corresponding labels of them are randomly selected from the dataset, as shown on left and right parts of the figure. The corresponding parts of these 6 images and labels are cropped and patched to form a new image and a new label, as shown by rectangles with transparent colors.

Experimental results in Section 4 show that the proposed method is able to improve segmentation accuracies of the deep neural net, and two main enhancements on the original RICAP can effectively boost the performance gain.

3.1. The enhanced random cropping and patching

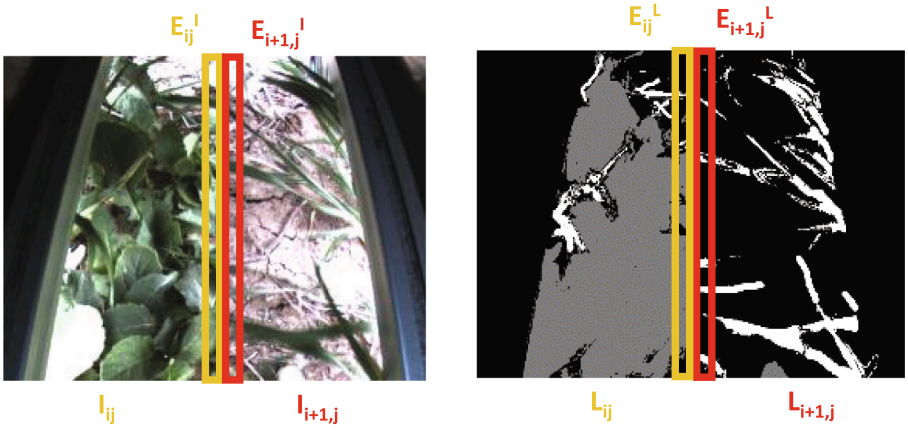
In the proposed framework, we enhance the conventional RICAP method by relaxing the fixed number of 4 parts cropping, and flexibly determining the number of crops according to the informative part of actual images. Firstly, we randomly crop the image region into H slices along the horizontal direction, and V slices along the vertical direction. Therefore there will be $H \times V$ number of independent regions. For each region, an image from the dataset is randomly selected and the corresponding region is patched to the new image. As an example shown in Fig. 1, for our image dataset recorded in a wheat farm using a wide angle camera inside a trailer, the image region is sliced into 2 parts horizontally and 3 parts vertically in this example, since the informative region of the image is a vertical rectangle region in the middle. The size of horizontal and vertical crops are completely random for each image to be generated.

Next, $H \times V$ number of images $I_{ij}(i = 1,\cdots,H, j = 1,\cdots,V)$, where i and j are horizontal and vertical index of each image crop region, and their corresponding labels are randomly selected from the dataset. These selected images and labels are cropped and patched together to form a new image and a new label as shown in Fig. 1. Specifically, the cropping and patching steps for region corresponding to I_{ij} can be formulated as follows,

$$\begin{aligned} w_i &= \text{round}(w_i' W_I) \\ h_j &= \text{round}(h_j' H_I) \end{aligned}, \tag{1}$$

where w_i and h_j are width and height of the crop region corresponding to I_{ij}; $\text{round}(\cdot)$ is the rounding to integer operation, and W_I and H_I are width and height of raw images in dataset; w_i' and h_j' are ratio of w_i and h_j of the crop region w.r.t. W_I and H_I of raw images, and can be computed as follows,

$$\begin{aligned} w_i' &= \frac{w_i^{rand}}{\sum_{i=1}^{H} w_i^{rand}} \\ h_j' &= \frac{h_j^{rand}}{\sum_{j=1}^{V} h_j^{rand}} \end{aligned}, \tag{2}$$

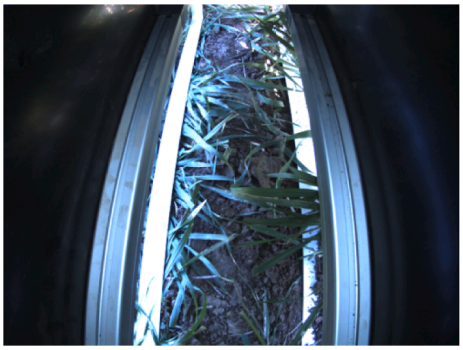


(a) Patched images I_{ij}, $I_{i+1,j}$ with their edge pixels E^I_{ij} and $E^I_{i+1,j}$.

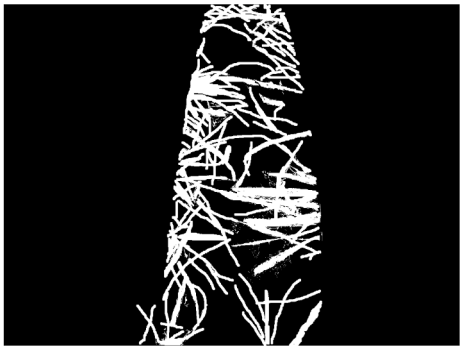
(b) Patched labels L_{ij}, $L_{i+1,j}$ with their edge pixels E^L_{ij} and $E^L_{i+1,j}$.

Fig. 2. Patching of two images I_{ij} and $I_{i+1,j}$, along with their corresponding labels L_{ij} and $L_{i+1,j}$. Difference in value of edge pixels E^I_{ij} and $E^I_{i+1,j}$ on colour images, and E^L_{ij} and $E^L_{i+1,j}$ on label images are used to compute the matching cost for these two images. In (b), grey pixels represent weed, while white pixels represent crop.

(a) The Digital Farmhand.



(b) Image 1.



(c) Label 1.

(d) Image 2.

(e) Label 2.

Fig. 3. The Digital Farmhand robot and typical images captured in the wheat farm. (c) and (e) are ground truth labels of (b) and (d) labeled manually. Images (b) and (d) are images captured in different weather conditions. Image (b) was captured in a cloudy day, while image (d) was captured in a sunny day which leads to over exposure on the left side of image.

in which w_i^{rand} and h_j^{rand} are randomly generated numbers correspond to $w_i^{'}$ and $h_j^{'}$, and they follow simple uniform distribution,

$$\begin{aligned} w_i^{rand} &= \mathscr{U}(0,1) \\ h_j^{rand} &= \mathscr{U}(0,1) \end{aligned}. \tag{3}$$

Finally, the top left and bottom right pixel indexes of the crop region correspond to I_{ij} are formulated as follows,

$$\begin{cases} x_{ij}^{tl} &= \sum_{k=1}^{i-1} w_k + 1 \\ y_{ij}^{tl} &= \sum_{k=1}^{j-1} h_k + 1 \\ x_{ij}^{br} &= \sum_{k=1}^{i} w_k \\ y_{ij}^{br} &= \sum_{k=1}^{j} h_k \end{cases}, \tag{4}$$

where $x_{ij}^{tl}, y_{ij}^{tl}, x_{ij}^{br}$, and y_{ij}^{br} are x, y coordinates along image width and height directions of top left and bottom right corners of the crop region corresponding to I_{ij}.

Although our proposed method is based on the original RICAP technique, there are various differences between them. On the one hand,

Table 1
Information of two datasets used in the paper before data augmentation.

Dataset	Bonn	Narrabri
Number of images	283	150
Resolution	1296 × 966	1288 × 964
Train, valid and test images	199,42,42	113,16,21
Percentage of background pixels	93.92%	85.80%
Percentage of weeds pixels	1.96%(veges weed)	7.56%(wheat weed)
Percentage of crop pixels	4.13%	6.64%

in our method, we keep the original size of the image without applying resizing operation as it is used in the original RICAP method. Keeping the original size of the cropped image enables the deep neural net to learn sizes of crop and weed leaves. This is important in our application, because images are recorded from a constant height of the camera from the ground for the agricultural robots we deploy. In comparison, the original RICAP is designed for generic image classification tasks, therefore it is critical for it to handle objects recorded from different camera perspectives. On the other hand, different from the original RICAP, class labels of $H \times V$ images are not fused for the semantic segmentation task, since each pixel is defined to be of one unique category and is treated independently, and cropping and patching have no effect on individual pixels and their labels. More discussions on segmentation accuracies for different configurations of image cropping will be presented in Section 4.2.

3.2. Multi-proposal generation and selection

For semantic segmentation tasks, individual pixel is assigned to a specific category not only based on its own RGB values, but also those of its neighboring pixels. Therefore, the cropping and patching operation produces a negative effect as it destroys connection of pixels along edges of each crop region to neighboring pixels. This effect gets worsened when edge pixels patched to new neighbors have distinctive colors and labels with their new neighboring pixels, as shown in Fig. 2. Along the boundaries of two colour images in Fig. 2(a) and corresponding labels in Fig. 2(b), there are striking differences since the left image mostly contains wild radish weed, while the right image mostly contains soil and wheat crop along the boundary.

To tackle such a negative effect, in our method, we introduce a multi-proposal generation and selection step to avoid generating images that have large differences along the boundaries of different crop regions. To be more specific, we introduce a matching cost to represent the mismatching of edge pixels of two patched images and their labels. As shown in Fig. 2, the edge pixels of left and right images (I_{ij} and $I_{i+1,j}$) and their labels (L_{ij} and $L_{i+1,j}$) are denoted as $E^I_{ij}, E^I_{i+1,j}, E^L_{ij}$ and $E^L_{i+1,j}$. Then, the matching cost between these two edge pixels is formulated as

$$\mathscr{M}_{ij}^{i+1,j} = (1-\alpha) \parallel E^I_{ij} - E^I_{i+1,j} \parallel^2 + \alpha \parallel E^L_{ij} - E^L_{i+1,j} \parallel^2, \tag{5}$$

where $\mathscr{M}_{ij}^{i+1,j}$ is the matching cost of the edge between cropped images I_{ij} and $I_{i+1,j}$, α is a weighting hyperparameter that is used to balance edge pixel differences between colour images and mask labels. In general, in order to ensure the matching costs from colour and label images have statistically similar orders of magnitude, the α in Eq. (5) can be determined by the following way,

$$\frac{\alpha}{1-\alpha} = \frac{\Sigma_{k=1}^{K} ||E^I_k - \check{E}^I_k||^2}{\Sigma_{k=1}^{K} ||E^L_k - \check{E}^L_k||^2}, \tag{6}$$

where E^I_k and $\check{E}^I_k$ denote kth edge pixel and its corresponding neighbouring edge pixel in colour image. E^L_k and $\check{E}^L_k$ denote kth edge pixel and its corresponding neighbouring edge pixel in label image. K is the total number of the edge pixels in a sample of a fixed number of patched augmentation images from the training data.

The matching cost of the whole generated image is sum of matching costs along all edges of crop regions:

$$\mathscr{M} = \sum_{i,j}^{H,V} \sum_{\gamma_i,\gamma_j \in \Gamma_{i,j}} \mathscr{M}_{ij}^{\gamma_i,\gamma_j}, \tag{7}$$

where Γ_{ij} represents neighboring cropped images of I_{ij}.

In the proposed method, P number of candidate patched images are proposed, and the one with the minimum matching cost is selected as the final generated image, i.e.,

$$p_{sel} = \underset{p}{\operatorname{argmin}} \lim \{\mathscr{M}_1, \cdots, \mathscr{M}_P\}, \tag{8}$$

where $\mathscr{M}_p$, for $p = 1, \cdots, P$, are the matching costs of the P proposals and p_{sel} is the index of the selected proposal. By employing such a multi-proposal generation and selection step, the negative effect caused by boundary mismatching of cropped images can be effectively minimised, and the performance of the trained deep neural nets is improved as a result, as elaborated in Section 4.3.

The whole procedure of the proposed data augmentation method for semantic segmentation is summarised in Algorithm 1.

Algorithm 1.

3.3. Datasets

The first dataset is collected by the authors of this paper in a wheat farm in Narrabri, NSW, Australia. A downward looking PointGrey BlackFly RGB camera is fixed in a trolley made of a metal box with a V shaped head. During the deployment stage, the trolley is dragged by the Digital Farmhand, a multi purpose agricultural robot developed by the Australian Centre for Field Robotics (ACFR), to move between two rows of wheat plants while the robot drives in a farm with wheat growing in different stages of their life cycles. The wheat plants are of 2, 4, and 6 weeks old, counting from their first emergence out of soil. Images captured from the wheat farm of different growth stages are incorporated into one dataset, to enrich the variety of the dataset. The dataset contains wheat of Feekes scale 2–7. The camera captures images under the canopy of the wheat plants between two rows of wheat continuously at 30 Hz. The captured images are used for semantic segmentation, which produces masks to classify each pixel in the image to be background, weed or crop. Finally, the mechanical and laser tools in the trolley execute weeding actions according to the generated masks. The robot and typical images captured in the wheat farm are shown in Fig. 3. The model of the camera is BFLY-U3-13S2C-CS, which is a global shutter camera and has a pixel size of 3.75μm. In order to capture images with larger spatial size, the camera is attached at a certain height. It is coupled with the lens of model LENS-28C2-V100CS, and the focal length is manually set to be as short as 3.0 mm, The benefit of camera being at this height and its focal length being short is that the captured image corresponds to a larger size of the ground, and two consecutive images have larger overlap so that the robot can drive faster. Images are captured in different weather conditions, which introduce different illumination conditions, as shown in Fig. 3(b) and Fig. 3(d). Images captured in sunny days are subject to over exposure on part of images as shown in Fig. 3(d). These different illumination conditions and over exposure make semantic segmentation problem very challenging. We labeled 150 labeled images out of over 10K captured images due to the time consuming labeling process. The labeled dataset is split into training, validation and testing by 75%–10%–15%. We refer it as "Narrabri" dataset in the following sections.

For the second dataset, we use the publicly available dataset (Chebrolu et al., 2017) captured in a sugar beet farm in Bonn, Germany. The images are captured with a 4-channel RGB + NIR camera JAI AD-130 GE, which is mounted in nadir view. Among these 4 channels, only the RGB channels are used in this paper, since the semantic segmentation net deployed here is designed for off-the-shelf RGB cameras. The Bonirob (Milioto et al., 2018) is deployed for the data collection. The Bonn dataset has near constant illumination, and none or very little overlapping between crop and weed since images are captured in early growth stage of crop, which makes semantic segmentation task simpler than the Narrabri dataset. For the Bonn dataset, it originally released 283 images with semantic labels. These images are used for data augmentation experiment here to demonstrate the performance of the proposed method, since the main purpose of this paper is to validate the proposed data augmentation framework, for the most common situation

Algorithm 1: Steps of the proposed data augmentation method for semantic segmentation.

Input:raw dataset with images and associated mask labels;
Output:A batch of N newly generated images with associated mask labels;
Created image number n = 0;
while $n < N$ **do**
  (1) create image region cropping template with H,V number of horizontal and vertical random slices using Eq. 1-Eq. 4;
  (2) create P proposals of generated images and labels;
  for p=1:P **do**
    (2-1) randomly select $H \times V$ images and their labels from raw dataset;
    (2-2) crop these selected images and labels from step (2-1) according to cropping template created in step (1), and patch them to generate a new proposal for the new image and label to be generated.;
    (2-3) compute matching cost for the generated proposal in step (2-2) using Eq. 5 and Eq. 7;
  end
  (3) select the proposal with the minimum matching cost among P proposals created in step (2) using Eq. 8;
  n++;
end

where only a limited number of labeled images are available. The dataset is split into training, validation and testing by 70%-15%-15%. Recently, the Bonn dataset has been extended to include around 10K labeled images, and these are only used in Section 4.4, to show the limitation of our method, *i.e.* the performance gain from data augmentation methods reduces when the number of training images increases.

The details about these two datasets before data augmentation is summarised in Table 1. Both Bonn dataset and Narrabri dataset are used to evaluate the proposed method thoroughly. Bonn dataset was collected from a sugar beet farm, and the captured images have near identical background illumination. On the other hand, images in Narrabri dataset were captured in a wheat farm, and have quite different background illuminations as can be seen from Fig. 3. There are challenging situations like over exposure exist in Narrabri dataset. By utilizing these two distinctively different datasets, the proposed data augmentation method can be evaluated more comprehensively.

3.4. The adopted deep neural net and training details

To evaluate the performance gain the proposed data augmentation method brings in, we use a state-of-the-art encode-decoder architecture based real time semantic segmentation deep neural net, Bonnet (Milioto and Stachniss, 2018) for the experiment. Bonnet allows for fast multi-GPU training, retraining, and deployment in a robotic system.

Bonnet architecture based on ERFNet was selected, with around 1.1M parameters and 9B operations. In order to truly reflect the quality of training dataset and not to be influenced by well optimised weights trained on several thousand images, the pre-trained weights are ignored and weights are randomly initialised in all experiments. The input image is 512×384.

During the training process, all data augmentation methods use the same number of training images and train the net with the same 20K iterations except the one explicitly set to be different in Table 3. By doing so, it ensures fair comparisons are made between different data augmentation methods. All other hyperparameters are set to be the same as those used by Bonnet for crop-weed segmentation. Specifically, median frequency is used to balance losses of different classes, gamma for focal loss is set to 2, and learning rate is set to 0.001 initially which decays by 1.5 every 10 epochs. Adam optimiser is used during training with its epsilon set to be $1e^{-8}$. All experiments are trained on a GTX1080Ti GPU with batch size of 16 images.

In all experimental evaluations, to balance the contributions of matching costs of image edges and label edges, during each trial of the same experiment, α in Eq. (5) is computed using a sample of 50 patched images and labels from the training data of that trial according to Eq. (6).

In the baseline configuration of our proposed method, we generate 10 proposals for each image, *i.e.* $P = 10$ in Eq. (8). We slice the image region into 5 vertically and 2 horizontally for images in Narrabri dataset as the informative part of images is mostly in the middle (see in Section 4.2 for more detailed discussions on the reason for this). However, we slice the image region in Bonn dataset into 4 both horizontally and vertically, as informative part of images in Bonn dataset expands the whole image region. The baseline configuration of the proposed method achieves the best performance as detailed below. We refer it as "baseline" in the remainder of the paper (see in Fig. 4, Fig. 5 and Table 2).

Table 2
The semantic segmentation accuracies of the deep neural net in test data of Narrabri and Bonn datasets with different data augmentations.

Dataset	Augmentation	mAcc(%)[mean(std,t)]	mIOU(%)[mean(std,t)]	Precision(%)[mean]			Recall(%)[mean]		
				background	weed	crop	background	weed	crop
Narrabri	Basic	91.01 (1.837, **4.996**)	63.59 (2.317, **7.863**)	96.433	71.893	52.175	94.382	87.105	50.705
	2 × 2 (Ours)	93.85 (0.391, **2.035**)	69.44 (1.701, **2.549**)	96.958	76.145	66.268	96.939	88.245	58.575
	Baseline (Ours)	**94.02** (0.414, -)	**70.77** (1.325, -)	**96.978**	**78.607**	**68.221**	**97.019**	**88.459**	**59.898**
Bonn	Basic	97.99 (0.338, **4.114**)	74.26 (0.981, **7.304**)	99.354	51.003	92.034	98.981	60.778	92.985
	2 × 2 (Ours)	98.49 (0.150, 0.453)	76.83 (0.515, **2.284**)	99.389	**58.113**	93.931	99.445	**63.598**	93.348
	Baseline (Ours)	**98.51** (0.107, -)	**77.09** (0.432, -)	**99.390**	57.991	**94.432**	**99.449**	63.590	**93.807**

Table 3
The semantic segmentation accuracies of the deep neural net in test data of Narrabri and Bonn datasets with no data augmentations, data augmentation of 2 × 2 patches without multi-proposal generation and selection, and the proposed baseline method.

Dataset	Augmentation	mAcc(%)[mean(std,t)]	mIOU(%)[mean(std,t)]	Precision(%)[mean]			Recall(%)[mean]		
				background	weed	crop	background	weed	crop
Narrabri	No	89.89 (0.920, **11.373**)	62.06 (1.901, **10.256**)	95.423	70.549	50.909	93.222	85.785	50.102
	No (More Iter)	89.90 (0.803, **11.771**)	62.11 (1.729, **10.860**)	95.451	70.550	51.118	93.235	85.847	50.075
	2 × 2(Without)	92.95 (0.384, **4.436**)	69.04 (1.105, **3.331**)	96.922	76.021	66.101	95.919	88.247	58.501
	Baseline (Ours)	**94.02** (0.414, -)	**70.77** (1.325, -)	**96.978**	**78.607**	**68.221**	**97.019**	**88.459**	**59.898**
Bonn	No	97.75 (0.187, **9.932**)	73.38 (0.482, **13.740**)	99.305	50.103	90.823	98.785	60.801	91.991
	No (More Iter)	97.75 (0.181, **9.726**)	73.44 (0.473, **13.479**)	99.311	50.101	91.001	98.777	60.818	92.012
	2 × 2(Without)	98.48 (0.144, 0.491)	76.89 (0.459, **1.989**)	99.385	**58.001**	94.127	99.438	**63.592**	93.452
	Baseline (Ours)	**98.51** (0.107, -)	**77.09** (0.432, -)	**99.390**	57.991	**94.432**	**99.449**	63.590	**93.807**

3.5. Evaluation metric

To compare performance of different data augmentations, we compute and show the pixel-wise performance of the deep neural net tested in the two datasets. In particular, we compute precisions and recalls of all classes, i.e., background, weed and crop, as well as mean accuracy and mean IOU (mIOU) of all pixels. The mean accuracy (mAcc) and the mean IOU are computed as follows:

$$mAcc = \frac{1}{C}\sum_{c=1}^{C} f_c \frac{TruePos_c}{TruePos_c + FalsePos_c}, \tag{9}$$

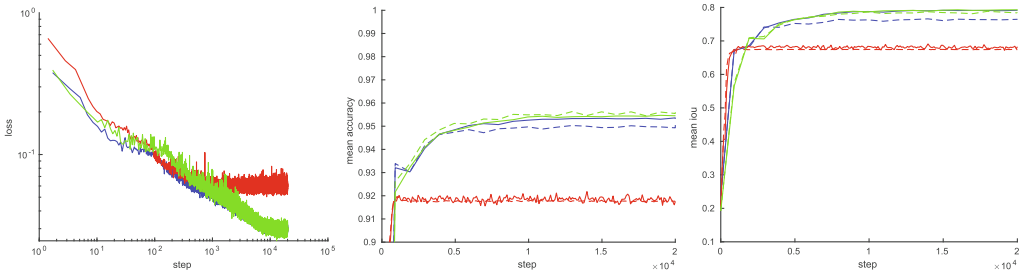
$$mIOU = \frac{1}{C}\sum_{c=1}^{C} \frac{TruePos_c}{TruePos_c + FalsePos_c + FalseNeg_c}, \tag{10}$$

where $TruePos_c, FalsePos_c, FalseNeg_c$ are true positively, false positively, and false negatively classified pixel numbers for cth class in the semantic segmentation problem. f_c is the pixel percentage of class c.

In order to evaluate different data augmentation methods thoroughly and make a fair comparison, statistical significance test between different algorithms have been conducted. In particular, the re-sampled paired t test (Dietterich, 1998) has been conducted to show the statistical significance between two different algorithms. To do that, we conduct 10 trials of train and test for each methods. During each trial, the available data is randomly divided into training, validation and testing data with the ratio mentioned in Section 3.3, and statistics (mean and Standard Deviation (STD)) of evaluation metrics in Eq. (9) and Eq. (10) of 10 trials are reported. Then a paired t test is used to determine the statistical significance of overall evaluation metrics of the proposed approach over others. Let's denote m_A^k and m_B^k to be the evaluation metrics of method A and B on kth trial. In addition, let's assume $m^k = m_A^k - m_B^k$ from 10 different trials are independently drawn from a normal distribution, a Student's t distribution can be applied by the following statistics (Dietterich, 1998),

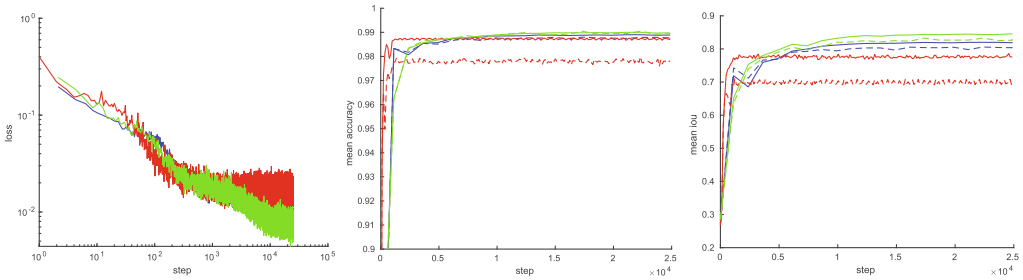
$$t = \frac{\overline{m}\sqrt{n}}{\sqrt{\frac{\sum_{i=1}^{n}(m^k - \overline{m})^2}{n-1}}}, \tag{11}$$

where $\overline{m} = \frac{1}{n}\sum_{i=1}^{n} m^k$ is the mean of m^k, and n is the total number of trials, which is 10 in our case. Under the null hypothesis, this statistic has t distribution of $n-1 = 9$ degree of freedom. The null hypothesis will be rejected, for 10 trials, if the computed $|t| > t_{9,95\%} = 1.833$, where $t_{9,95\%}$ means 9 degrees of freedom and the significance level of 95%. In other word, if the mean value of the evaluation metrics of the proposed



(a) Loss over optimisation steps. (b) Mean accuracies over optimisation steps. (c) Mean IOU over optimisation steps.

Fig. 4. Reduction of loss (a), and mean accuracies and mean IOU (b)(c) over optimisation steps with different data augmentations of Narrabri dataset. The blue, red and green solid lines in (a), (b), (c) represent our proposed baseline approach, the basic data augmentation and the proposed approach of 2 × 2 patches on training data. The dash lines in (b), (c) represent mean accuracies and mean IOU in validation data.



(a) Loss over optimisation steps. (b) Mean accuracies over optimisation steps. (c) Mean IOU over optimisation steps.

Fig. 5. Reduction of loss (a), and mean accuracies and mean IOU (b)(c) over optimisation steps with different data augmentations of Bonn dataset. The blue, red and green solid lines in (a), (b), (c) represent our proposed baseline approach, the basic data augmentation and the proposed approach of 2×2 patches on training data. The dash lines in (b), (c) represent mean accuracies and mean IOU in validation data.

baseline method outperforms the other method and the computed $|t| > 1.833$, the proposed method is supposed to yield statistically significant superior performance than the other method.

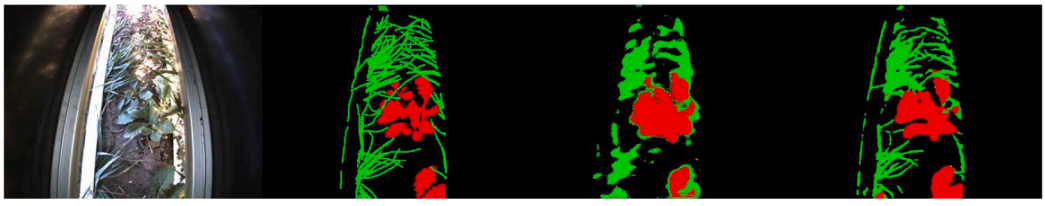
4. Experimental results

4.1. Comparisons against conventional methods

In this section, we compare the semantic segmentation accuracies of the deep neural net in Section 3.4 with three different ways of augmenting the two datasets detailed in Section 3.3. The deep neural net is trained with a basic data augmentation, with the proposed data augmentation of 2×2 patches like the original RICAP (but with multi-proposal generation and selection as in Section 3.2), and with our proposed data augmentation with the baseline configuration.

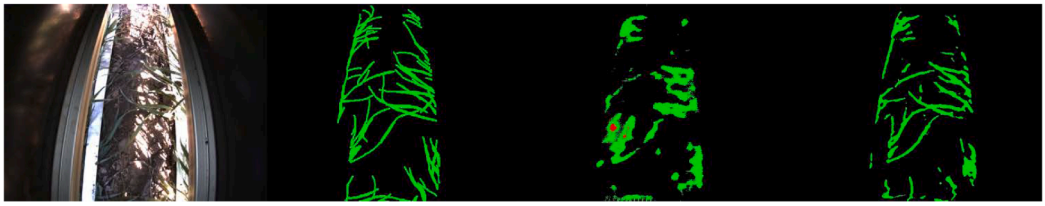
The basic data augmentation uses a conventional way of augmenting data by randomly flipping images horizontally and vertically, shifting the gamma value of the images and blurring the images. We refer it as "basic" later in the paper (see in Fig. 4, Fig. 5 and Table 2). To compare the proposed method with the original RICAP with the proposed multi-proposal generation and selection, images are sliced into 2×2 patches, which we refer as "2×2" in Fig. 4, Fig. 5 and Table 2.

For three different data augmentation methods using the Narrabri and Bonn datasets, the performance comparisons over optimisation steps of one particular trial of the experiment are shown in Fig. 4 and Fig. 5 as an typical example. In both figures, the blue, red and green solid lines in (a), (b), (c) represent reduction of loss, mean accuracies and mean IOU with our proposed baseline approach, the basic data augmentation and the proposed approach of 2×2 patches on training data, respectively. The dash lines in (b), (c) represent mean accuracies



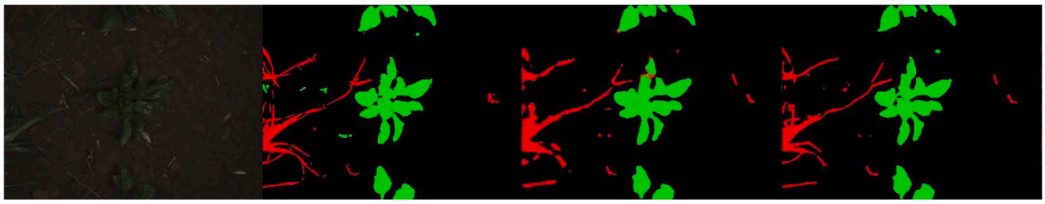
(a) Example segmentation result 1.

(b) Example segmentation result 2.

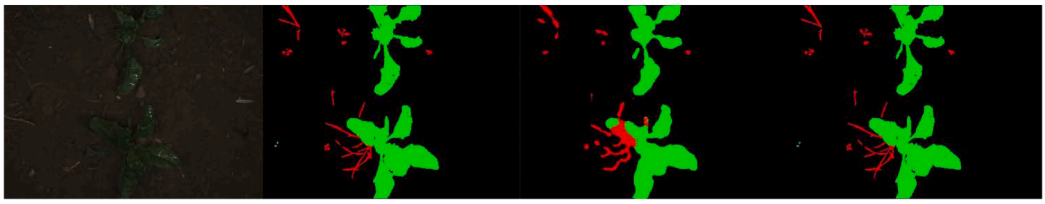


(c) Example segmentation result 3.

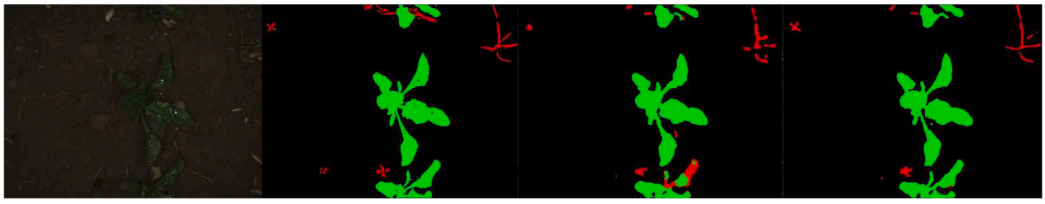
Fig. 6. Qualitative comparison of segmentation results with the basic data augmentation and the proposed baseline method in Narrabri dataset. In each subfigure, images from left to right are the raw RGB image, the ground truth segmentation mask, and predicted segmentation masks from the deep neural net with the basic data augmentation and the proposed baseline method of data augmentation.



(a) Example segmentation result 1.



(b) Example segmentation result 2.



(c) Example segmentation result 3.

Fig. 7. Qualitative comparison of segmentation results with the basic data augmentation and the proposed baseline method in Bonn dataset. In each subfigure, images from left to right are the raw RGB image, the ground truth segmentation mask, and predicted segmentation masks from the deep neural net with the basic data augmentation and the proposed baseline method of data augmentation.

and mean IOU in validation data. From Fig. 4(a) and Fig. 5 (a), we can see that the optimisation has converged after more than 10^4 steps. Note that all horizontal and vertical axes of (a) in both figures are log scaled. For both datasets, the minimisation of losses of the proposed baseline method and the proposed method with 2×2 patches show similar performance, and both of them are better than that of the basic data augmentation technique. Similar situations are found for mean accuracies and mean IOU for both datasets, as shown in Fig. 4 (b)(c) and Fig. 5(b)(c).

The segmentation accuracies in test data are detailed in Table 2. The bold numbers corresponding to the mean value of evaluation metrics in the table show the best performing method. The bold numbers corresponding to the paired t test values (short noted as t in the table) of mean accuracy and mean IOU denote the proposed baseline approach is statistically significant than the other method, *i.e.* $|t| > t_{9,95\%} = 1.833$. It can be seen that the proposed data augmentation method with the baseline configuration shows the best performance in both mean accuracies and mean IOU in both datasets. The proposed method with 2×2 patches shows better performance than the basic data augmentation scenario. These results confirm that the proposed method can effectively improve performance of a deep neural net for semantic segmentation by increasing variety of training images from relatively small number of labeled images, which is a common situation in agricultural datasets. It can also be argued that, the 2×2 patches used in the original RICAP method, together with the proposed multi-proposal generation and selection step in Section 3.2, can be enhanced to any $H \times V$ horizontal and vertical patches that suit images in the dataset best, as it is used in the proposed method with the baseline configuration. Further insights on how H and V can be determined are provided in Section 4.2.

For the basic data augmentation and the proposed baseline data augmentation using the Narrabri and Bonn datasets, the qualitative comparisons of semantic segmentation results, are shown in Fig. 6 and Fig. 7, respectively. As can be seen from Fig. 6, the segmentation masks of the basic data augmentation are generally blobby and fine details of crop and weed leaves are missing. In addition, part of crop leaf in Fig. 6 (c) is misclassified as weed. Similar situation can be found in the Bonn dataset. Weed leaves appear to be more blobby and some parts of crop leaves in Fig. 7(b) and Fig. 7(c) are misclassified as weed. In comparison, the segmentation results with our proposed baseline method achieve better accuracy, and finer details of crop and weed leaves are captured. Segmentation results from the proposed method with 2×2 patches are visually similar to those of the baseline method, therefore they are not shown here. These results show that the proposed data augmentation method can effectively improve performance of a deep neural net for semantic segmentation.

In terms of the operational timing, the proposed data augmentation

Table 4
The semantic segmentation accuracies of the deep neural net w.r.t. random or uniform image region cropping in test data of Narrabri and Bonn datasets.

Dataset	Augmentation	mAcc(%)[mean(std,t)]	mIOU(%)[mean(std,t)]	Precision(%)[mean]			Recall(%)[mean]		
				background	weed	crop	background	weed	crop
Narrabri	Uniform	93.78 (0.420, **2.158**)	69.38 (1.183, **3.402**)	96.971	75.163	66.589	96.780	87.979	59.870
	Baseline (random)	**94.02** (0.414, -)	**70.77** (1.325, -)	**96.978**	**78.607**	**68.221**	**97.019**	**88.459**	**59.898**
Bonn	Uniform	98.35 (0.311, **2.008**)	76.21 (0.390, **4.751**)	**99.409**	55.025	94.087	99.305	63.012	93.520
	Baseline (random)	**98.51** (0.107, -)	**77.09** (0.432, -)	99.390	**57.991**	**94.432**	**99.449**	**63.590**	**93.807**

Table 5
The semantic segmentation accuracies of the deep neural net w.r.t. different number of image patches in test data of Narrabri and Bonn datasets.

Dataset	Augmentation	mAcc(%)[mean(std,t)]	mIOU(%)[mean(std,t)]	Precision(%)[mean]			Recall(%)[mean]		
				background	weed	crop	background	weed	crop
Narrabri	2×2	93.85 (0.391, **2.035**)	69.44 (1.701, **2.549**)	96.958	76.145	66.268	96.939	88.245	58.575
	2×4	93.37 (0.409, **2.936**)	69.55 (1.390, **2.287**)	96.919	75.553	67.232	96.294	**90.195**	58.289
	Baseline (2×5)	**94.02** (0.414, -)	**70.77** (1.325, -)	96.978	**78.607**	68.221	**97.019**	88.459	59.898
	2×6	93.95 (0.370, 0.862)	70.12 (1.192, 1.315)	96.952	77.788	67.035	97.003	87.253	60.091
	2×8	93.72 (0.511, 1.720)	70.06 (1.390, **1.851**)	**97.011**	75.812	**68.374**	96.633	89.187	59.901
	2×10	93.12 (0.342, **5.950**)	69.13 (1.875, **3.289**)	96.941	75.999	65.880	96.051	87.020	**60.231**
Bonn	2×2	98.49 (0.150, 0.453)	76.83 (0.515, **2.284**)	99.389	**58.113**	93.931	99.445	63.598	93.348
	Baseline (4×4)	**98.51** (0.107, -)	**77.09** (0.432, -)	**99.390**	57.991	**94.432**	**99.449**	63.590	**93.807**
	8×8	98.27 (0.148, **3.381**)	75.46 (0.823, **4.525**)	99.200	51.340	94.014	99.191	**65.807**	92.589
	10×10	98.22 (0.211, **4.872**)	75.38 (0.925, **5.483**)	99.289	52.897	93.835	99.207	64.298	91.990

method with baseline configuration averagely takes around 10 ms with parallel processing, therefore does not add too much additional time during training.

In addition, we also compare the proposed baseline approach with no data augmentation, no data augmentation and with $50K$ optimization steps, and data augmentation of 2×2 patches without multi-proposal generation and selection step. The results are shown in Table 3, where "No (More Iter)" denotes no data augmentation and with $50K$ optimization steps, and "2×2(Without)" denotes data augmentation of 2×2 patches without the multi-proposal generation and selection step. It can be seen that the segmentation accuracies decrease when no data augmentation is used, and training more iterations does not help increasing the performance. The proposed data augmentation method with the baseline configuration shows clearly better performance than no augmentation cases, and better overall performance than 2×2 patches without the multi-proposal generation and selection step.

4.2. Properties of the proposed method

In this section, we investigate various properties of the proposed data augmentation method. Specifically, we analyse the influence of random or uniform image region cropping and different number of image region cropping on the segmentation accuracies of the deep neural net.

Firstly, in the baseline implementation of the proposed method, the size of each cropped region of the image is determined randomly, as given in Eq. (1) to Eq. (4). A comparison of the proposed method with uniform cropping and random cropping is made here to check if random cropping can boost accuracies of segmentation results. The results of segmentation accuracies in test data of Narrabri and Bonn datasets are shown in Table 4. It can be seen from Table 4 that the random cropping yields better overall performance than the uniform cropping. Our explanation is that the generated images have more variability when being cropped randomly. When the augmented dataset with more variability is used for training, it prevents the model from overfitting the dataset and improves the segmentation accuracies in test data.

Secondly, we further investigate the influence of number of crop regions on the segmentation accuracies. It has been shown in Section 4.1 and Table 2 that the 2×2 patches used in the original RICAP method can be generalised to any $H \times V$ horizontal and vertical patches that suit images in the dataset best. We further investigate how different number of crop regions and direction of cropping influence the performance of the segmentation net. As such, a comparison of segmentation accuracies of the deep neural net trained with different number of $2, 4, 5$ (baseline), $6, 8$ and 10 vertical patches V is carried out for the Narrabri data. For the Bonn dataset, a comparison of segmentation accuracies trained with 2×2, 4×4 (baseline), 8×8 and 10×10 image patches is carried out. The results are shown in Table 5. The results show that increasing number of vertical patches V from 2 to 5 improves the segmentation accuracies for Narrabri dataset in general, however further increasing it not only doesn't help but also decrease the performance. One intuitive explanation is that the size of the informative region of images from the Narrabri dataset has horizontal and vertical aspect ratio close to $2:5$; therefore the deep neural net trained with such patches achieves the best accuracies. Larger number of patches, *e.g.* 2×10 patches in Narrabri dataset and 10×10 patches in Bonn dataset, make each patch region too small to capture a whole size leaf of crop or weed, thereby degrading the performance of the trained neural net.

Finally, we further investigate another interesting property of the proposed method. In the proposed method, each patched image part comes from the same location of the candidate image, as shown in Fig. 1. For example, the upper left part of the patched image comes from the same upper left part of the candidate image. Here we examine what if the patched image does not come from the same location of the candidate, but any random location of the candidate image. The results are shown in Table 6, and we denote the way of patching image from any random location of the candidate image as "scramble" in the table. It can be seen from the results that the proposed baseline approach yields better results in Narrabri dataset, whereas the two ways of patching yield similar performance (neither of them yields statistically significant mean accuracy or mean IOU than the other) in Bonn dataset. Our explanation is that the images from Narrabri dataset have clear structure due to the existence of the machine wall at left and right sides of the image. Therefore, patching images from the same location of the candidate image maintains such a structure in the image, and makes the patched image better resemble the real image. In contrast, the image in the Bonn dataset does not have such a structure, and therefore two ways of patching yield similar performance.

4.3. Ablation study

As explained earlier, we make two novel enhancements to the original RICAP method, namely, (1) relaxation of the fixed number of 2×2

Table 6
The semantic segmentation accuracies of the deep neural net w.r.t. different patching approaches in test data of Narrabri and Bonn datasets.

Dataset	Augmentation	mAcc(%)[mean(std,t)]	mIOU(%)[mean(std,t)]	Precision(%)[mean]			Recall(%)[mean]		
				background	weed	crop	background	weed	crop
Narrabri	Scramble	93.74 (0.490, **2.234**)	70.15 (1.431, **1.857**)	96.975	77.002	67.891	96.708	**88.470**	59.758
	Baseline	**94.02** (0.414, -)	**70.77** (1.325, -)	**96.978**	**78.607**	**68.221**	**97.019**	88.459	**59.898**
Bonn	Scramble	98.50 (0.119, 0.216)	77.01 (0.398, 0.532)	99.385	**58.011**	**94.441**	**99.451**	63.528	93.553
	Baseline	**98.51** (0.107, -)	**77.09** (0.432, -)	**99.390**	57.991	94.432	99.449	**63.590**	**93.807**

Table 7
The semantic segmentation accuracies of the deep neural net w.r.t. different multi proposals generation and selection strategies in test data of Narrabri and Bonn datasets.

Dataset	Augmentation	mAcc(%)[mean(std,t)]	mIOU(%)[mean(std,t)]	Precision(%)[mean]			Recall(%)[mean]		
				background	weed	crop	background	weed	crop
Narrabri	Without	93.61 (0.452, **2.610**)	69.12 (1.890, **3.300**)	96.969	75.357	66.548	96.707	88.278	58.135
	Image	93.78 (0.401, **1.846**)	70.03 (1.239, 1.380)	**97.001**	77.277	67.383	96.801	87.698	59.798
	Label	93.96 (0.372, 0.558)	70.26 (1.198, 1.001)	96.989	77.595	67.893	**97.021**	**88.519**	58.954
	Baseline (both)	**94.02** (0.414, -)	**70.77** (1.325, -)	96.978	**78.607**	**68.221**	97.019	88.459	**59.898**
Bonn	Without	98.36 (0.311, **2.023**)	75.78 (0.702, **4.471**)	99.370	53.750	93.909	99.334	62.895	93.109
	Image	98.43 (0.118, **2.202**)	76.23 (0.498, **3.308**)	99.381	54.573	94.109	99.378	**63.701**	93.458
	Label	98.47 (0.123, 1.293)	76.41 (0.456, **2.637**)	**99.403**	55.389	94.377	99.430	63.508	93.289
	Baseline (both)	**98.51** (0.107, -)	**77.09** (0.432, -)	99.390	**57.991**	**94.432**	**99.449**	63.590	**93.807**

Table 8
The semantic segmentation accuracies of the deep neural net w.r.t. different data augmentation methods in test data of full or small part of Bonn dataset.

Dataset	Augmentation	mAcc(%)[mean(std,t)]	mIOU(%)[mean(std,t)]	Precision(%)[mean]			Recall(%)[mean]		
				background	weed	crop	background	weed	crop
Bonn	Basic (full)	**98.55** (0.091, 1.130)	77.35 (0.431, **2.222**)	99.395	58.030	**94.459**	**99.458**	**64.012**	94.395
	Baseline (full)	**98.55** (0.095, 1.003)	**77.38** (0.391, **2.294**)	**99.401**	**58.103**	94.437	99.451	63.910	**94.512**
	Baseline (part)	98.51 (0.107, -)	77.09 (0.432, -)	99.390	57.991	94.432	99.449	63.590	93.807

patches, and (2) multi-proposal generation for each image and selection of the best proposal based on the defined matching cost in Eq. (7) and Eq. (8). The first enhancement has already been validated in details by results in Section 4.1 and Section 4.2. In this section, we evaluate the effect of the newly introduced multi-proposal generation and selection in improving the performance of the trained deep neural net for semantic segmentation tasks. In particular, we compare the segmentation accuracies trained without multi-proposal generation and selection step, and multi-proposal generation and selection with only cropped image edges are considered (*i.e.*, $\alpha = 0$ in Eq. (5), only cropped label edges are considered (*i.e.*, $\alpha = 1$ in Eq. (5), and the baseline of both cropped image edges and label edges are considered (*i.e.*, α in Eq. (5) is computed using Eq. (6), for Narrabri and Bonn datasets.

The results are shown in Table 7. It can be seen from Table 7 that the proposed method with the baseline configuration achieves superior overall performance with statistical significance than the scenario without multi-proposal generation and selection step in both datasets. Compared to the proposed methods with multi-proposal generation and selection considering only cropped image edges or cropped label edges, the proposed baseline approach also mostly achieves better overall accuracies in Bonn dataset, and better or similar performance in Narrabri dataset. These results show that the introduced multi-proposal generation and selection step is indeed crucial in improving accuracies of the trained deep neural net for semantic segmentation.

4.4. Limitation

In this Section, we examine the performance of the proposed data augmentation method when more than 10,000 labeled images are available for training. To do that, we add additional data from the newly released Bonn extended dataset to the training data of each trial of the experiments. We train the model with this "full" data using basic (random flipping, rotation and colour jitters, denoted as "Basic (full)") and the proposed data augmentation (denoted as "Baseline (full)"), and compare it against the baseline approach with original Bonn dataset of 283 labeled images (denoted as "Baseline (part)"). The results are shown in Table 8. We can see from the results that both approaches using full data yields better mean IOU with statistical significance than the baseline approach with less training data. The performances of basic and the proposed data augmentation methods using the full data yield similar performance without statistical significance (the paired t test value between the two methods are less than 0.001 in terms of mean accuracy and 0.182 in terms of mean IOU). The results indicate that the performance gain brought by the proposed data augmentation method degrades when a large number of training images are available.

5. Conclusions

In this paper, we have presented a novel data augmentation framework for image semantic segmentation in ground robotic applications in agriculture. More specifically, we borrow concepts from the conventional RICAP data augmentation method originally designed for data augmentation in image classification, and propose two novel enhancements to the former framework so that it can be used effectively for crop and weed semantic segmentation task. Comparison to the conventional methods, comprehensive evaluation of different properties of the proposed method and ablation studies have been carried out. The results show that the proposed method can effectively improve the segmentation accuracies of the deep neural net and the introduced enhancements over the original RICAP are essential for the obtained performance gain. Finally, we remark that the performance gain brought by the proposed method degrades when a large number, *e.g.* around 10,000, of training data is available as a limitation of the proposed method. The potential future works include real time detecting of crop and weed with the robotic platform, and execution of the weeding action based on the detected weed plants.

CRediT authorship contribution statement

Daobilige Su: Conceptualization, Methodology, Software, Writing – original draft. **He Kong:** Conceptualization, Methodology, Writing – review & editing, Supervision. **Yongliang Qiao:** Software, Writing – review & editing. **Salah Sukkarieh:** Writing – review & editing, Supervision.

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.

References

Bah, M., Hafiane, A., Canals, R., 2018. Deep learning with unsupervised data labeling for weed detection in line crops in UAV images. Remote Sensing 11, 1690.

Bosilj, P., Aptoula, E., Duckett, T., Cielniak, G., 2019. Transfer learning between crop types for semantic segmentation of crops versus weeds in precision agriculture. J. Field Robot.

Cap, Q.H., Uga, H., Kagiwada, S., Iyatomi, H., 2020. Leafgan: An effective data augmentation method for practical plant disease diagnosis. IEEE Trans. Autom. Sci. Eng..

Chebrolu, N., Lottes, P., Schaefer, A., Winterhalter, W., Burgard, W., Stachniss, C., 2017. Agricultural robot dataset for plant classification, localization and mapping on sugar beet fields. Int. J. Robot. Res. 36 (10), 1045–1052.

Chlingaryan, A., Sukkarieh, S., Whelan, B., 2018. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. Comput. Electron. Agric. 151, 61–69.

Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V., 2018. Autoaugment: Learning augmentation policies from data. In: arXiv preprint arXiv:1805.09501.

DeVries, T., Taylor, G.W., 2017. Improved regularization of convolutional neural networks with cutout. In: arXiv preprint arXiv:1708.04552.

Di Cicco, M., Potena, C., Grisetti, G., Pretto, A., 2017. Automatic model based dataset generation for fast and accurate crop and weeds detection. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2017). pp. 5188–5195.

Dietterich, T.G., 1998. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput. 10 (7), 1895–1923.

Facebook, A.I., 2019. Reimplementation of ResNet by Facebook AI. https://github.com/facebook/fb.resnet.torch, [Online; accessed 19-March-2019].

Goodfellow, I.J., Shlens, J., Szegedy, C., 2014. Explaining and harnessing adversarial examples. In: arXiv preprint arXiv:1412.6572.

Hall, D., Dayoub, F., Kulk, J., McCool, C., 2017. Towards unsupervised weed scouting for agricultural robotics. In: 2017 IEEE International Conference on Robotics and Automation (ICRA 2017). pp. 5223–5230.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Identity mappings in deep residual networks. In: 2016 European Conference on Computer Vision (ECCV 2016). pp. 630–645.

Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R., 2012. Improving neural networks by preventing co-adaptation of feature detectors. In: arXiv preprint arXiv:1207.0580.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (NIPS2012). pp. 1097–1105.

Liu, B., Tan, C., Li, S., He, J., Wang, H., 2020. A data augmentation method based on generative adversarial networks for grape leaf disease identification. IEEE Access 8, 102188–102198.

McCool, C., Beattie, J., Firn, J., Lehnert, C., Kulk, J., Bawden, O., Russell, R., Perez, T., 2018. Efficacy of mechanical weeding tools: a study into alternative weed management strategies enabled by robotics. IEEE Robot. Automat. Lett. 3 (2), 1184–1190.

Milioto, A., Lottes, P., Stachniss, C., 2018. Real-time semantic segmentation of crop and weed for precision agriculture robots leveraging background knowledge in cnns. In: 2018 IEEE International Conference on Robotics and Automation (ICRA 2018). pp. 2229–2235.

Milioto, A., Stachniss, C., 2018. Bonnet: An open-source training and deployment framework for semantic segmentation in robotics using CNNs. In: arXiv preprint arXiv:1802.08960.

Minervini, M., Fischbach, A., Scharr, H., Tsaftaris, S.A., 2016. Finely-grained annotated datasets for image-based plant phenotyping. Pattern Recognit. Lett. 81, 80–89.

Mortensen, A.K., Bender, A., Whelan, B., Barbour, M.M., Sukkarieh, S., Karstoft, H., Gislum, R., 2018. Segmentation of lettuce in coloured 3D point clouds for fresh weight estimation. Comput. Electron. Agric. 154, 373–381.

Potena, C., Nardi, D., Pretto, A., 2016. Fast and accurate crop and weed identification with summarized train sets for precision agriculture. In: 2016 International Conference on Intelligent Autonomous Systems, pp. 105–121.

Sa, I., Chen, Z., Popovic, M., Khanna, R., Liebisch, F., Nieto, J., Siegwart, R., 2018. weedNet: Dense semantic weed classification using multispectral images and MAV for smart farming. IEEE Robot. Automat. Lett. 3 (1), 588–595.

Takahashi, R., Matsubara, T., Uehara, K., 2018. Data augmentation using random image cropping and patching for deep CNNs. In: arXiv preprint arXiv:1811.09030 (also appear in proc. of the 10th Asian Conference on Machine Learning, 2018).

Su, Daobilige, Qiao, Yongliang, Kong, He, Sukkarieh, Salah, 2021. Real time detection of inter-row ryegrass in wheat farms using deep learning. Biosystems Engineering 204, 198–211. https://doi.org/10.1016/j.biosystemseng.2021.01.019. In this issue.

Wang, Y., Huang, G., Song, S., Pan, X., Xia, Y., Wu, C., 2021. Regularizing deep networks with semantic data augmentation. IEEE Trans. Pattern Anal. Mach. Intell.

Zagoruyko, S., Komodakis, N., 2016. Wide residual networks. In: arXiv preprint arXiv:1605.07146.

Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D., 2017. mixup: Beyond empirical risk minimization. In: arXiv preprint arXiv:1710.09412.

Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y., 2020. Random erasing data augmentation. In: AAAI Conference on Artificial Intelligence, pp. 13001–13008.

Zoph, B., Cubuk, E.D., Ghiasi, G., Lin, T.-Y., Shlens, J., Le, Q.V., 2020. Learning data augmentation strategies for object detection. In: European Conference on Computer Vision (ECCV 2020). pp. 566–583.