Transfer learning between crop types for semantic segmentation of crops versus weeds in precision agriculture

Petra Bosilj
School of Computer Science
University of Lincoln
Brayford Way, Brayford Pool
LN6 7TS Lincoln, UK
pbosilj@lincoln.ac.uk

Erchan Aptoula
Institute of Information Technologies
Gebze Technical University
41400, Kocaeli, Turkey
eaptoula@gtu.edu.tr

Tom Duckett
School of Computer Science
University of Lincoln
Brayford Way, Brayford Pool
LN6 7TS Lincoln, UK
tduckett@lincoln.ac.uk

Grzegorz Cielniak
School of Computer Science
University of Lincoln
Brayford Way, Brayford Pool
LN6 7TS Lincoln, UK
gcielniak@lincoln.ac.uk

Abstract

Agricultural robots rely on semantic segmentation for distinguishing between crops and weeds in order to perform selective treatments, increase yield and crop health while reducing the amount of chemicals used. Deep learning approaches have recently achieved both excellent classification performance and real-time execution. However, these techniques also rely on a large amount of training data, requiring a substantial labelling effort, both of which are scarce in precision agriculture. Additional design efforts are required to achieve commercially viable performance levels under varying environmental conditions and crop growth stages. In this paper, we explore the role of knowledge transfer between deep-learning-based classifiers for different crop types, with the goal of reducing the retraining time and labelling efforts required for a new crop. We examine the classification performance on three datasets with different crop types and containing a variety of weeds, and compare the performance and retraining efforts required when using data labelled at pixel level with partially labelled data obtained through a less time-consuming procedure of annotating the segmentation output. We show that transfer learning between different crop types is possible, and reduces training times for up to 80%. Furthermore, we show that even when the data used for re-training is imperfectly annotated, the classification performance is within 2% of that of networks trained with laboriously annotated pixel-precision data.

1 Introduction

In addition to increasing crop health and yield, a key objective in new and sustainable farming methods is to reduce the reliance on herbicides and pesticides, in order to decrease their negative environmental side-effects such as food and soil contamination, and reduction of biodiversity. Vision-guided autonomous

systems are being developed in precision agriculture to monitor the key indicators of crop status and decide on the correct treatment. Scene analysis allows such systems to apply targeted and per-plant treatments instead of treating the whole field uniformly, utilising procedures such as mechanical weed removal and spot spraying with herbicides, pesticides or fertilisers.

Such vision systems typically include a plant classification component to identify individual crop and weed plants, which is then used as a basis for selecting the required treatment. A number of different vision pipelines, typically consisting of vegetation segmentation followed by classification and based on hand-crafted features, have been proposed (Hemming and Rath, 2001; Lottes et al., 2017; Haug et al., 2014; Kusumam et al., 2017; Bosilj et al., 2018a). Conversely, Convolutional Neural Networks (CNNs) can automatically determine complex and highly discriminative features directly from images. Recently developed CNN-based weed and crop classification systems (Mortensen et al., 2016; McCool et al., 2017; Milioto et al., 2018; Sa et al., 2018; Lottes et al., 2018) also integrate background removal (e.g. separating plants from soil) into the classification step, and have been reported to achieve excellent classification performance as well as real-time processing speeds, a strong requirement for the the commercial deployment of such systems. One of the remaining obstacles towards the commercial feasibility of weed and crop classification systems is the high annotation effort required to generate the amount of labelled data needed for the training, as well as for adaptation to different environmental and crop conditions (Slaughter et al., 2008; Lottes et al., 2018). Different approaches to alleviating this drawback are being actively explored (Di Cicco et al., 2017; Hall et al., 2017; Lottes et al., 2018).

In this work we explore the transfer of knowledge between CNNs trained to work with different crop types with the goal of reducing the annotation and training effort needed for the deployment of crop-weed classification systems on a variety of crop types, field conditions and weed species. Our goal is to reduce the training efforts typically required for a new crop, by reusing the networks trained of different crop types. To this end, we work with a publicly available dataset for sugar beets (Chebrolu et al., 2017), as well as our own datasets representing carrot and onion crops.

The contributions presented in this paper are as follows:

- we publish two annotated datasets showing carrots and onions respectively, consisting of RGB and NIR information and annotated ground truth images[1],
- we demonstrate that by transferring the knowledge from a network trained on a different crop, training times can be reduced to less than 20% of those needed to train a CNN classifier from scratch,
- we show that a CNN classifier can be successfully fine-tuned using partially and imprecisely labelled images obtained with minimal annotation effort.

In the next section, we give a brief overview of related work, focusing primarily on precision agriculture systems relying on CNNs. The classifier used in this work is presented in Sec. 3, followed by the explanation of our experimental setup and the evaluation results in Sec. 4. The paper is concluded in Sec. 5.

2 Related work

Initial attempts at plant type identification in agriculture were typically implemented as two-step systems, consisting of a background (soil) removal step, followed by classification of the remaining vegetation areas. The background segmentation can be done directly on a colour image in various colour spaces (Ruiz-Ruiz et al., 2009), or on index-images calculated from different spectra such as Normalized Difference Vegetation

[1]Available at `https://lcas.lincoln.ac.uk/wp/research/data-sets-software/crop-vs-weed-discrimination-dataset/`.

Index (NDVI) (Rouse Jr et al., 1974), using global (Otsu, 1979), local (Bosilj et al., 2018b), or machine learning approaches (Guerrero et al., 2012), where increased precision typically comes at a cost of decreased speed due to the computational complexity of the approaches. The foreground corresponding to vegetation is then presented to a classifier, working either with connected regions or at pixel or grid precision. The advantage of a region-based approach is the speed of classifying only a few samples per image, however they can not cope with regions of mixed vegetation making them unsuitable to use at the later growth stages of most crops, where crops and weeds start to overlap and pixel-based classifiers become mandatory (Lottes et al., 2017). The samples (either regions or pixels) are described by hand-crafted discriminative features, presented to the classifier together with the ground truth label in the training stage, and used later for online classification.

In parallel to the aforementioned developments, the advent of deep learning in its various forms has led to a paradigm shift in pattern recognition (Lecun et al., 1998), motivated by ground-breaking performance increases in many challenging computer vision tasks. The capacity of deep networks to automatically learn a hierarchy of increasingly complex features from their input data has practically eliminated the process of hand-crafted feature construction. Consequently, it is no surprise that deep-learning-based techniques are being adapted rapidly for application in precision agriculture. For example, in (Mortensen et al., 2016), a deep CNN is trained to classify different types of crops in order to estimate their biomass. (Potena et al., 2016) process RGB and NIR data with a cascade of two CNNs, where the first network performs the background removal step, while the second one further classifies the remaining vegetation into crops and weeds. The approach adopted by (McCool et al., 2017) is based on training a complex network and then replacing it with an ensemble of lightweight CNNs trained from the original network.

More recently, real-time systems to distinguish crops from weeds have been developed (Sa et al., 2018; Milioto et al., 2018; Lottes et al., 2018) based on SegNet (Badrinarayanan et al., 2017), a deep encoder-decoder convolutional network architecture for multi-class pixel-wise segmentation. The performance of the classifiers trained on different combinations of input channels as well as on NDVI images calculated from the Red and NIR channels is compared in (Sa et al., 2018), without a significant difference between the different combinations. They also studied the impact of using a pre-trained (on generic color images) network model followed by fine-tuning, however they concluded that this did not have a significant impact on the output. A classifier relying only on RGB data was shown to be competitive with a system trained on RGB and NIR data, if the RGB data can be supplemented by various alternate data representations (Milioto et al., 2018).

While these systems have demonstrated the suitability of SegNet for crop-weed classification, and achieved excellent precision and execution speed on the target crop and field, obtaining the required training data remains a challenge. In (Sa et al., 2018), the training data was specifically collected for the project, by cultivating a pure crop, pure weed and a testing plot through controlling the herbicide dosage. Both the time required to grow such plots, as well as the annotation time reported for the testing plot containing both crop and weed species makes this data collection set-up problematic for commercial deployment. In (Milioto et al., 2018), they investigate the number of images needed to re-train a classifier for a new field, however they target the same crop at different growth stages and require precise annotations. Without labelled images from the new target field, the approach presented in (Milioto et al., 2018) reaches acceptable performance on new fields only with the addition of a sequential module exploiting spatial crop arrangement (Lottes et al., 2018), but at the expense of real-time processing speed. It was shown in (Di Cicco et al., 2017) that a deep learning system can be trained successfully with synthetically generated training data. However, generating this data requires information about the leaf distribution in different plant species as well as high definition RGB leaf textures. A different approach was presented by (Hall et al., 2017), where the training regions were first clustered to minimize the number of regions that the domain expert needs to process in order to reduce the effort required in the data annotation process.

Since it is not realistic to train a generic CNN classifier that would work on all crop and weed types, we instead focus on the need to adapt the existing systems rapidly to a new crop type. We compare the time and performance of retrained classifiers to those trained from scratch for a certain crop type. We also study the ability of these networks to adapt to a new crop type when the annotations are obtained with much more

time-efficient region-based labelling, which however produces images where labels are missing for patches of mixed vegetation and increases the chance of mislabelling during the annotation process.

3 Approach

Unlike the aforementioned existing work, our goal is not classification performance maximization. Instead, we have focused on exploring whether a neural network trained with images for a given crop can be *efficiently* adapted to a different crop type without significant loss of segmentation performance. Hence, we have chosen to use the same fundamental SegNet architecture employed in various forms in recent related work (Sa et al., 2018; Milioto et al., 2018; Lottes et al., 2018). In more detail, our classification pipeline is an end-to-end system consisting of a fully connected CNN, which receives as input a stack of colour and near-infrared images, extracts their pixel features, and then performs pixel-level classification to output a segmentation map with three classes: soil, crops and weeds (Fig. 1).

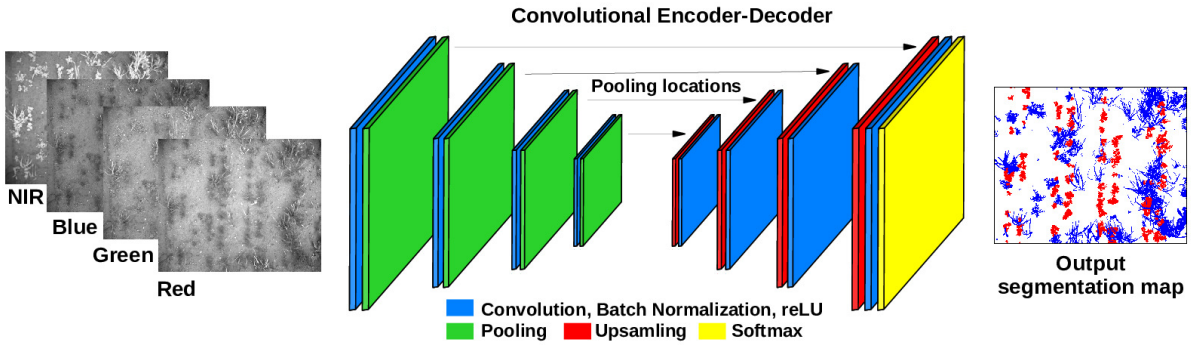

Figure 1: Outline of the employed SegNet-Basic architecture.

Input: The network input comprises RGB colour and near-infrared (NIR) images of size 512 × 384 pixels, stacked together to form a three-dimensional tensor of size 4 × 384 × 512. The input is not preprocessed or augmented in any way. This was a conscious decision, with the end of designing an approach as close to commercial deployment conditions as possible, We plan to further investigate a possibility of a a generalised preprocessing pipeline or data augmentation technique which would be optimal and useful independently of the crop type and the conditions in which the images are acquired. Moreover, conversely to general purpose semantic segmentation where commonly hundreds of thousands of training samples are available, this is a context where data acquisition and labelling is highly cost-intensive, so regardless of the specifics of the method, it must be able to cope with the scarcity of training data.

Segmentation framework: Our segmentation framework is a simplified version of the well-known SegNet architecture (Badrinarayanan et al., 2017), called SegNet-Basic. Segnet has been designed specifically for efficient general-purpose pixel-level semantic segmentation of images and video. It is known to be capable of handling hundreds of classes and is commonly trained with hundreds of thousands of samples. However, since our domain is characterised by severely limited dataset sizes, only 3 classes and a high operational speed requirement, it was deemed necessary to simplify the aforementioned network from 13 layers to 4 in order to both decrease the number of involved parameters and the duration of training and testing. Our decision is also supported by the similar observations made in past papers (Sa et al., 2018; Milioto et al., 2018; Lottes et al., 2018) about how relatively simple networks are sufficient for dealing with this problem.

More specifically, SegNet is based on an "encoder-decoder" architecture for content description, followed by a pixel-wise multi-class softmax classifier. Encoder-decoders are a type of neural network specialising in

learning effective feature representations. They task the encoder part of the network with mapping, layer by layer, its raw input to progressively more compact and abstract feature representations. The decoder part of the network then admits this feature representation as input, processes it further in order to perform a decision and produces the final network output.

From a practical perspective, the encoder applies convolutions to the input and progressively down-samples it through max-pooling to produce an effective feature representation. It is then followed by the decoder that then up-samples this representation, leading to dense feature maps, which are finally classified through a multi-class softmax classifier.

In more detail, each encoding layer of our network consists of a convolution with 64 filters of constant size 7×7 that are tasked with learning an increasingly abstract hierarchy of image features, layer by layer. The filter size controls the size of the pixel neighbourhood that will contribute to the final decision on the central pixel's label. If it is too small, the decision will be too localised, and miss useful surrounding information. On the other hand, if it is too large, it risks associating the pixel under study with neighbouring distinct classes. The chosen filter size is the one recommended by the designers of SegNet designers so as to achieve optimal performance.

Convolutions are followed by batch normalisation, a well-known technique for improving the stability and performance of neural networks. Next, the rectified linear unit (reLU) activation function is applied to enable the network to learn non-linear features, followed finally by the max-pooling stage that down-samples the feature map to half of its original size.

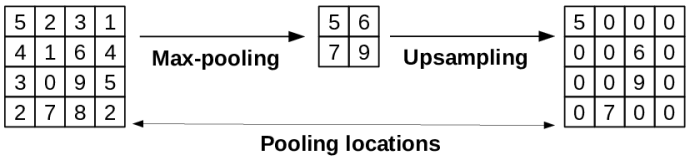

Figure 2: Example of the max-pooling and upsampling technique used in SegNet.

After the fourth encoding layer, the same number of decoder layers begin. Each decoding layer first convolves its input with the same settings as in encoding convolutions, and then applies batch normalisation. The feature map is then upsampled to double its original size. Upsampling is conducted through placing the pixel values at their original locations, safeguarded previously during encoding, and filling the rest of the pixel locations with zeroes (Fig. 2). This leads to sparse feature maps, which are then densified with the following decoding layer's convolution step.

Class weighting: Conversely to scene labelling with CNNs, one of the reasons why semantic segmentation is more challenging is the highly unbalanced distribution of classes within the same image. Evidently, soil pixels dominate any given field scene with respect to either crop or weed. Consequently, the cross-entropy loss function needs to be tuned carefully, with class specific weights, such that any false predictions concerning weeds and/or crops will be penalised more higly with respect to soil. To this end, we employ median frequency weighting (Eigen and Fergus, 2015), where each class is weighted by $w_c = \tilde{f}/f(c)$, with $f(c)$ representing the number of pixels of class c divided by the total number of pixels in the training set, and $\tilde{f}$ is the median of the three class frequencies.

Output: The last layer of the network consists of a softmax activation function, which produces a real vector of length 3 per image pixel. Each dimension of this vector represents the probability of the pixel under consideration belonging to the soil, crop or weed class, respectively.

4 Experimental set-up and data

4.1 Dataset characteristics

To validate our approach, we use three different sources of multi-spectral vegetation images representing different crops (Table 1). We use the provided RGB colour information and the NIR channel directly, rather than combining them into NDVI (Rouse Jr et al., 1974), as often done in vegetation detection and plant recognition. We made this choice to simplify the classification inputs, motivated by the fact that (Sa et al., 2018) report minimal difference in performance when including NDVI information directly. This is most likely due to the ability of the network to learn how to calculate the NDVI representation from the given inputs.

The **Sugar Beets 2016 (SB16)** is a public dataset depicting a broad-leaf crop sown in a single crop row, providing RGBN information collected under controlled lighting conditions with the BoniRob agricultural robot (Chebrolu et al., 2017). The dataset was collected using the JAI 130-GE camera, which uses a splitting prism resulting in perfectly aligned image channels. The opaque shroud in which the camera was mounted provides independence from natural lighting. The original image size is 1296×966 pixels with spatial resolution of 3 px/mm depicting a field patch of dimensions 24×31 cm. This dataset exhibits the least amount of weed pressure, and the majority of both crop and weed plants belong to distinct connected regions in the image. The dataset provides pixel-based ground truth for 280 images, distinguishing between the background (mostly soil), sugar beets and several types of weeds. For the purpose of this experiment, all the different weed types were treated as one class. Example images from this dataset are shown in Fig. 3(a) and 3(d).

The **Carrots 2017 (CA17)** dataset was collected by the authors in the fields of Lincolnshire, UK, in June 2017 using two Teledyne DALSA Genie Nano cameras providing RGB and NIR information respectively, mounted 5 cm apart. This produced high resolution images of size 2464×2056 corresponding to a patch of ground of approximately 100×85 cm at a spatial resolution of 2.5 px/mm. No lighting control mechanisms were employed, and the camera set-up was mounted on a manually pulled cart. Since the optical centres of the cameras were not aligned, a registration step is required, resulting in the loss of several pixels at the sides of the aligned images, which additionally exhibit parallax errors, especially on taller vegetation. The images show a field under substantial weed pressure and contain weeds with a similar appearance to the crop, especially compared to *SB16*. Several regions of vegetation contain both crops and weeds in close proximity. Pixel-based ground truth is provided for 20 aligned images. Figs. 3(b) and 3(e) depict example images from this dataset.

The last dataset used is **Onions 2017 (ON17)**, which was also collected by the authors in the fields of Lincolnshire, UK, in April 2017 with the same acquisition set-up as *CA17*. The datasets thus share characteristics such as image and spatial resolution, as well as parallax errors noticeable in the registered images. The amount of weed pressure on this dataset falls between *CA17* and *SB16*, and occasional regions of mixed vegetation are present. The ground truth is provided for 20 annotated images. Example images can be seen in Fig. 3(c) and 3(f).

Annotations: Due to the substantial time efforts required to provide pixel labels for high-resolution images, the annotation process for the two datasets collected by the authors involved two steps. Firstly, the regions outputted by an existing vegetation segmentation approach (Bosilj et al., 2018b) were annotated as either crop, weed or mixed regions (with certain over-segmented and duplicate regions discarded). Then, these annotations were corrected by manually labelling the regions of mixed vegetation as well as removing any remaining over-segmentations around the edges of the vegetation regions. The annotation of *CA17* images, showing substantial amounts of vegetation, took around 20 minutes for the region annotation and an additional 3–4 hours for pixel-precision labelling. The images of *ON17* took about 15 minutes and 2 hours per image for region and pixel-precision labelling, respectively, due to a comparatively simpler scene. Both types of annotations are used in our experiments, providing five distinct sets of data together with SB16:

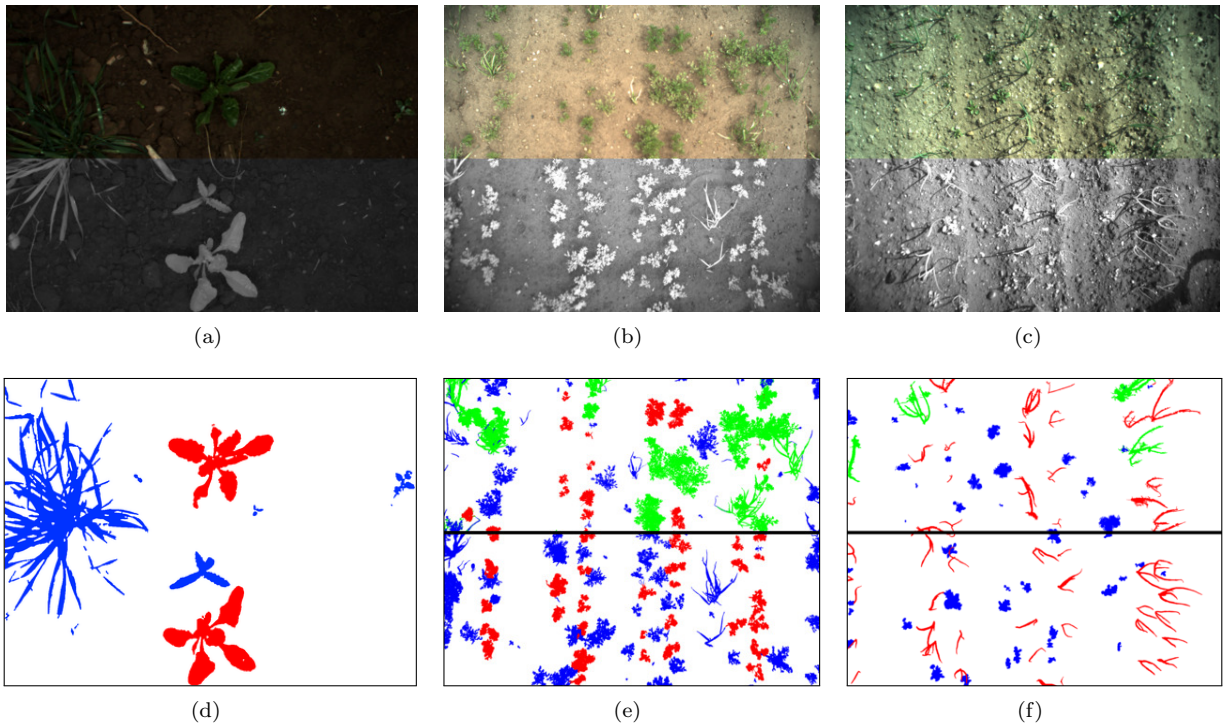

Figure 3: Examples of input images and ground truth for the three datasets. The input images are shown in (a) for SB16, (b) for CA17 and (c) for ON17, with the top half of the image showing the RGB image and the bottom the NIR channel. The ground truth for the corresponding images is shown in (d) for SB16, (e) for CA17 and (f) for ON17, where the images for CA17 and ON17 are showing the partial ground truth in the top half and the fully annotated images in the bottom half. In all of the ground truth images, the crops are shown in red, the weeds in blue, soil in white, and the mixed class in partially annotated images is shown in green.

- *fully labelled data*: **SB16**, **CA17-full**, **ON17-full**; comprises images coupled with their pixel-precision labelled ground truth,

- *partially labelled data*: **CA17-partial**, **ON17-partial**; comprises images coupled with region labelled images, where some regions are marked as mixed.

All three datasets compared in Table 1 show crops of very different visual appearance from one another. The weed pressure ranges from the strongest in *CA17* to lowest in *SB16*. While the crop growth stage in *SB16* and *ON17* is early enough that most of the crop plants can be clearly distinguished from one another, the crop in *CA17* is in later growth stages. There are cases where the carrot tops of 2–3 plants in a row merge into one vegetation region. The later growth stages of both crops and weeds in *CA17* are also evident from the vegetation coverage per image, which is 3–3.5 times that in *SB16* and *ON17*. Comparing the amount of data shown, it can be observed that the images from both *CA17* and *ON17* show an 11 times larger surface area at a similar spatial resolution compared to *SB16*. A similar observation can be made on the amount of plants shown per image in different datasets, where, compared to *SB16*, *CA17* has 40 times more crop plants and 10 times more weed plants per image while *ON17* has 23 times more crops and 3 times more weed plants. From this we conclude that the amount of content shown in the modest amount of images of *CA17* and *ON17* datasets is comparable to that shown across all images of *SB16*.

Table 1: Details of the datasets used in our experiments. In the class distribution column 'C' stands for 'crop' and 'W' for 'weeds'.

Dataset	#img	image size	vegetation share	class distrib.	crop rows	plants /img (avg)	area /img (cm^2)	annotations
Sugar Beets 2016	283	1296 × 966	6% of image	C - 68% W - 32%	1	C - 2.25 W - 8	24 × 31	pixel-precision
Carrots 2017	20	2428 × 1985	21% of image	C - 35% W - 65%	4 × 2	C - 88 W - 85.6	100 × 85	labelled regions & pixel-precision
Onions 2017	20	2419 × 1986	7% of image	C - 51% W - 49%	4 × 2	C - 52.3 W - 22.3	100 × 85	labelled regions & pixel-precision

4.2 Experimental set-up

Our experiments were designed to study the benefits of transfer learning between networks trained on different crops. The goal is to compare a fully trained classifier to a classifier retrained under realistic conditions in terms of training time, annotation effort required as well as classification performance. We perform three sets of experiments, for all datasets where the appropriate data is available:

1. *Train on crop X, test on crop X, using fully labelled data.* This is to obtain the best theoretical performance for each crop, with a network trained under controlled laboratory settings with no time limitations. These experiments provide the baseline network performance and training time for each crop.

2. *Train on crop X, retrain and test on crop Y, using fully labelled data.* This is to obtain the upper performance limit when transferring the knowledge from one trained network to another, resulting in faster network training. This still does not correspond to a realistic case for field deployment, as pixel-based annotation required to produce full ground truth is extremely time demanding.

3. *Train on crop X, retrain and test on crop Y, using the partially labelled data for retraining.* This set-up explores the viability of deploying a re-trained network on a field with a previously unknown crop. The images used are annotated only on a region basis, requiring little annotation effort. The network ignores all the input pixels belonging to mixed vegetation clusters, which are not considered during training. Despite a slight expected drop in performance, both the annotation effort and retraining are fast enough for field deployment of the system.

We process all the images at their original scale. In order to obtain the samples of the resolution required by the chosen network architecture, we divide the input images into samples of the appropriate size, starting from the top-left image corner, and discard any extra image pixels. For *SB16*, this means obtaining $2 \times 2 = 4$ samples from each of the input images, while for *CA17* and *ON17* we obtain $4 \times 4 = 16$ samples per image. Thus, all of the trained networks expect input samples of the same size, which facilitates transfer learning between datasets containing images of different sizes.

The networks under study were all trained using the stochastic gradient descent method. As our datasets are of relatively limited size when compared to general purpose computer vision tasks, and since due to practical reasons long term training is of no interest to this context, we used a fixed learning rate of 0.01 and the highest batch size that our hardware allowed, i.e. 4. Furthermore, during the transfer experiments (i.e. sets 2 and 3), in order to motivate the network to focus on retraining the weights of the classifier (i.e. final layer),

we have increased the learning rate for that layer ten-fold higher (resulting in the learning rate 0.1 for the final layer in these experiments).

During the first set of experiments, all networks were trained up to 50000 iterations (approximately 200 epochs), while we only trained up to 10000 iterations in the case of transfer networks (experiment sets 2 and 3). All datasets were split into a training set comprising of 80% of available images, while the remaining 20% were used as the test set. The images were sampled after dividing them into the training and testing set, such that no samples coming from the same image are present in both training and testing. All CNN implementations were based on the Caffe framework (Jia et al., 2014). The experiments have been conducted using a desktop computer equipped with an Intel i7 CPU, 32GB of memory and a GTX 1080Ti GPU with 11GB of video memory, where every 1000 training iterations lasted approximately 12 minutes.

4.3 Evaluation metrics

In order to evaluate and analyse the performance of different trained classifiers, as well as enable the comparison with related work, several performance measures were used. The first set of measures refer to the pixel-wise performance of the classifier. Instead of accuracy (Acc) which is a poor indicator of classification quality when the class distribution is very unbalanced, like in all the used datasets, we chose to express the performance through Cohen's Kappa coefficient (Cohen, 1960) which takes expected probability of a chance agreement between the classifier and the ground truth $\mathrm{Acc}_{\mathrm{chance}}$ into account:

$$\kappa = \frac{\mathrm{Acc} - \mathrm{Acc}_{\mathrm{chance}}}{1 - \mathrm{Acc}_{\mathrm{chance}}}. \tag{1}$$

We also calculate the precision and recall metrics for each class to examine the performance in more detail. Precision indicated the probability of a correct prediction when classifying a sample into a certain class c:

$$p_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c}, \tag{2}$$

while recall indicates the probability of correctly classifying a sample belonging to a class c:

$$r = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{3}$$

where TP_c, FP_c and FN_c are the number of true positive, false positive and false negative samples per class, respectively.

As the target agricultural applications such as precision weeding or spraying ultimately require the information describing individual plants, we additionally calculate the object-wise performance of each classifier in terms of precision and recall as well. This metric, proposed by the authors of *SB16* dataset (Milioto et al., 2018), additionally allows a better comparison between the presented work and other work done on this dataset. In order to calculate TP_c, FP_c and FN_c, for each connected component of the weed and crop class from the ground truth, the class-wise contributions of the predictions to this object are calculated. A positive example is given when the majority of the predictions belong to the label c, otherwise it is counted as a negative example. Precision and recall are then calculated following Eq. (2) and (3). Note that the metric is more suitable for the datasets where most of the vegetation belongs to distinct regions in the ground truth, such as *ON17* and *SB16*, as a single region of a class c can represent multiple plants of the same class when the vegetation in the dataset is in the later stages of growth like in *CA17*.

4.4 Results and discussion

The evaluation results of all the three sets of experiments are summarized in Table 2. In the first set of experiments, when the networks were directly trained with the annotated field images of the target crop, we

Table 2: Evaluation results. 'train' denotes the dataset from which the training set was constructed, 'weights' denotes which of the three networks trained from scratch was used for fine-tuning, while 'test' denotes the dataset from which the test set was taken.

Data				Pixel-based							Object-based			
train	weights	test	iter. (× 1000)	Soil		Weed		Crop		κ	Weed		Crop	
				p	r	p	r	p	r		p	r	p	r
Train on crop X, test on crop X, with fully labelled data														
SB16	–	SB16	45	99.91	98.99	66.05	94.48	94.71	97.46	91.24	95.53	82.21	84.42	73.03
CA17-f	–	CA17-f	28	98.16	96.38	80.63	87.02	75.97	77.68	83.24	89.95	80.00	90.58	89.78
ON17-f	–	ON17-f	39	99.62	98.72	83.76	92.79	72.28	86.64	84.88	92.96	88.00	95.76	93.39
Train on crop X, retrain and test on crop Y, with fully labelled data														
SB16	CA17-f	SB16	9.7	99.94	98.58	59.67	95.58	92.29	97.31	88.74	95.69	85.42	85.45	78.33
SB16	ON17-f	SB16	7.4	99.93	98.28	52.92	96.24	92.33	95.60	86.42	98.40	88.46	92.68	82.01
CA17-f	SB16	CA17-f	5.5	97.81	96.58	81.97	85.12	75.29	79.56	83.05	87.86	77.02	89.08	88.70
CA17-f	ON17-f	CA17-f	5.9	98.15	96.26	81.03	86.51	74.27	79.07	83.05	91.04	82.13	91.00	88.89
ON17-f	SB16	ON17-f	9.0	99.62	98.65	82.44	92.22	71.39	86.43	84.21	95.77	90.67	97.37	93.28
ON17-f	CA17-f	ON17-f	6.9	99.51	98.62	89.31	87.59	65.80	89.24	83.26	94.20	86.67	96.61	94.21
Train on crop X, retrain and test on crop Y, using partially labelled data for retraining														
CA17-p	SB16	CA17-f	0.9	98.04	94.82	78.18	82.08	65.39	83.43	79.37	80.69	69.36	84.27	82.61
CA17-p	ON17-f	CA17-f	2.7	97.94	94.87	76.91	81.39	67.06	82.87	79.04	82.23	68.94	85.66	83.94
ON17-p	SB16	ON17-f	10.0	99.67	98.45	80.30	93.94	70.15	86.91	83.52	94.52	92.00	96.52	93.28
ON17-p	CA17-f	ON17-f	9.3	99.69	98.29	77.28	94.90	70.15	86.62	82.66	91.78	89.33	94.92	92.56

achieve excellent agreement to the ground, with $\kappa = 91\%$ for *SB16* dataset, $\kappa = 83\%$ when training on *CA17* and $\kappa = 85\%$ with *ON17* dataset. The performances for each class on the *SB16* dataset are comparable with the latest results through a similar classification system (Lottes et al., 2018). Even though the only previously published results on the *CA17* are based on classifying only the vegetation regions (Bosilj et al., 2018a) and ignoring the soil class, a vast improvement in both precision and recall as well as κ (κ increased from 33% to 79–83%) indicates a much better performance of the network classifiers presented here. A similar observation can be made when comparing the performance of the network-based classifier trained on *ON17* to that of a region-based classifier (Bosilj et al., 2018b) (the object-based metrics for crops have improved for 5–15%, and for weeds for 5–10%). However, as expected when training the networks from scratch, the best performance is achieved late in the training (iterations 28000–45000, depending on the dataset), which takes a substantial amount of time (approximately 6 - 9 hours respectively).

The next set of experiments is designed to test the feasibility of retraining a network originally trained to classify one crop for classifying a different type of crop. The obtained results are typical for the case where network weights are initialised from a network trained on a similar computer vision task (Ghazi et al., 2017). In other words it achieves a very similar performance much faster. On the *SB16* dataset, this means a decrease in κ of less than 3%, while the object-wise performance improved over all the measures. The performance of the networks retrained with *CA17* and *ON17* data both decrease for less than 1% in terms of κ, and exhibit very similar object-wise performance to their counterparts from the first set of experiments. The training time also reduced to around 20% of time needed in the first set of experiments, or between approximately 60 and 100 minutes depending on the dataset. We have additionally retrained a publicly available model of SegNet-Basic, originally trained on the CamVid dataset (Brostow et al., 2009) for semantic segmentation of videos captured from a driving automobile, with *CA17* data. While the convergence time (i.e. the number of iterations required to reach peak performance) is similar to fine-tuning networks trained on agricultural data, the resulting performance of $\kappa = 72.51\%$ is significantly lower than retraining our models. We were, however, limited in this experiment with the lack of available pre-trained models for SegNet-Basic as well as public datasets containing NIR data. As the CamVid dataset did not contain NIR data, it was also omitted during retraining with *CA17*, contributing to the drop in performance. We can conclude that the available pre-trained out-of-domain models have limited use for applications in the agricultural domain.

Finally, in the last set of experiments we examine the behaviours of the networks re-trained in the same manner, but this time by presenting the network only with the examples of partially labelled data. The performance drop in terms of κ when using the *ON17-partial* data of around 1% is comparable to the one

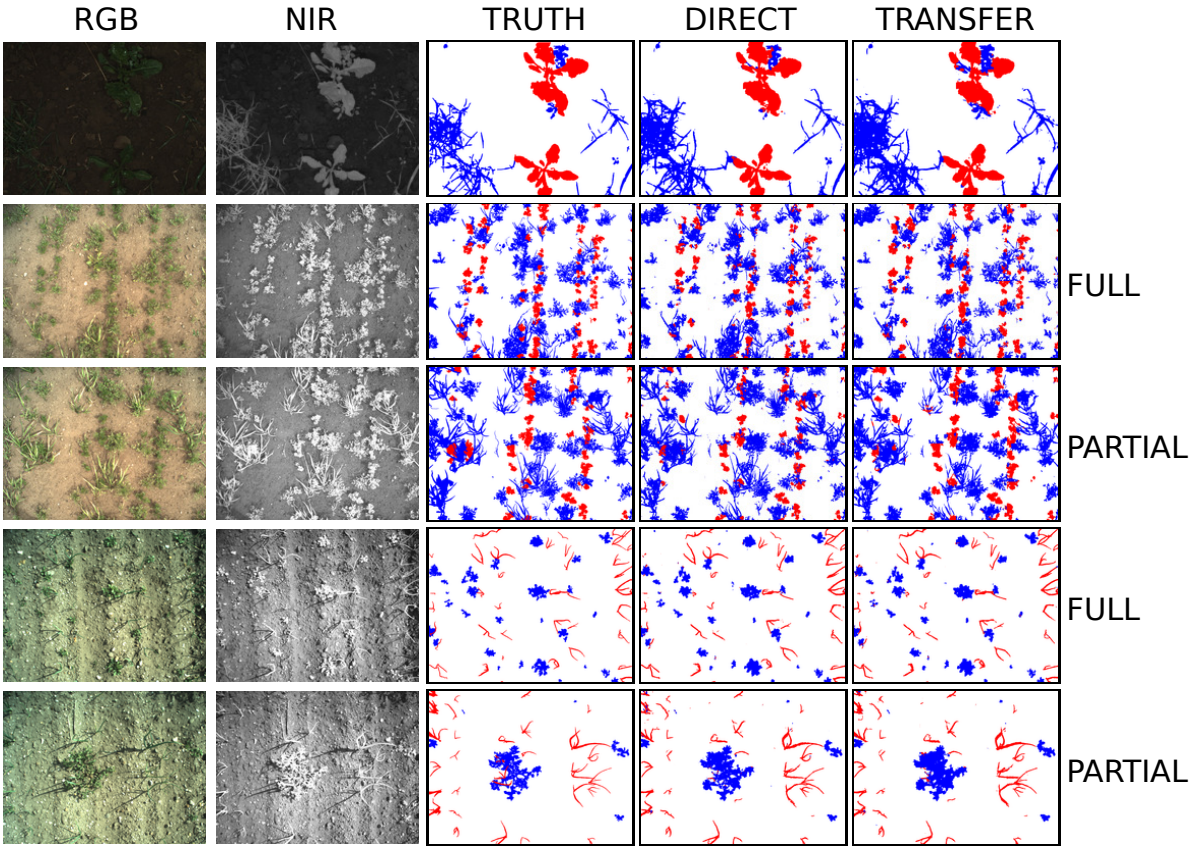

Figure 4: Examples of network inputs and outputs for different network configurations. The first two columns correspond to inputs of the network and the third one to the ground truth. The fourth column is the output of the classifier directly trained for the shown crop, while the fifth column shows the performance best transfer learning classifier (in case of *CA17* and *ON17* chosen between all the trained models of both transfer learning experiments). For the last three columns, red pixels correspond to crops, blue to weeds, and white to the soil class.

when using the *ON17-full* data, while the performance drop when using *CA17-partial* is larger at around 4%. This is likely due to the *CA17* dataset showing a field under a substantial weed pressure and thus containing substantially more regions of mixed vegetation for which the labels are not provided in the training process. Despite this, networks trained for both crops tested in this set of experiments still perform very well ($\kappa = 79\%$ on *CA17* dataset and $\kappa = 83\%$ on *ON17* dataset), while the training times are reduced by a similar factor to when the networks are retrained on fully annotated data (between approximately 10 to 90 minutes depending on the dataset).

We calculated the Bayesian confidence intervals for all the reported pixel-based measures, computed as the standard deviation of the Beta distribution posteriors over the belief in the accuracy assuming flat priors, confirming the significance of the results up to two decimal places. To further confirm the significance of the results, we have selected one representative classifier from each set of experiments for k-fold cross validation (with $k = 5$), and calculated the median κ value and 95% confidence intervals (which correspond to 1.96σ). We selected the classifiers targeting the *CA17* dataset, and obtained $\kappa = 82.96 \pm 4.0\%$ for training from scratch, $\kappa = 82.57 \pm 3.9\%$ when retraining the network trained on *SB16* with *CA17-full* and $\kappa = 79.24 \pm 3.2\%$ when retraining the same network with *CA17-partial*. A very small difference between the κ values obtained through k-fold cross validation and the ones reported in Table 2 confirms the validity of the reported performance.

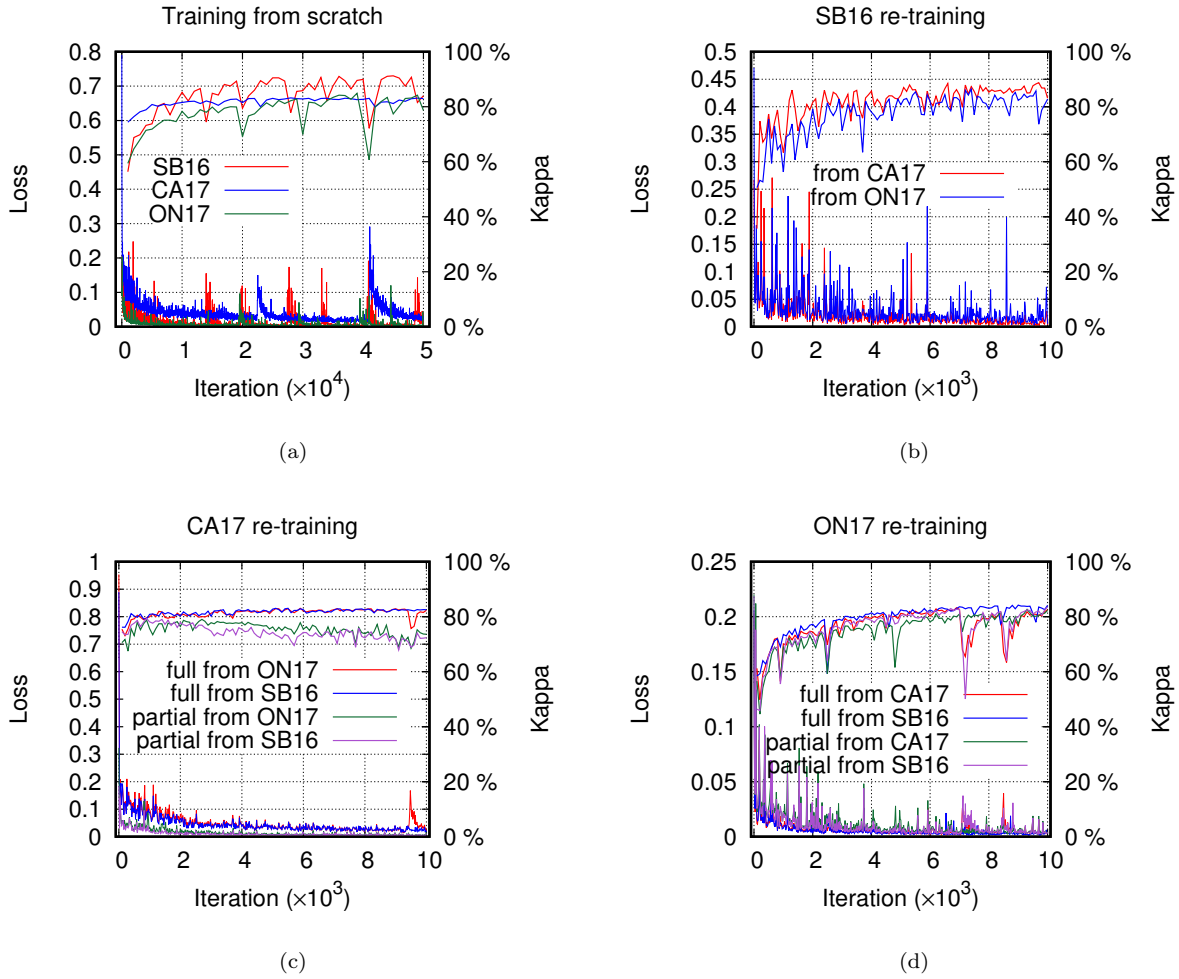

Figure 5: Loss per iteration during network training as calculated on the training set, and kappa per iteration calculated on the test set for each of the trained networks. The curves on the bottom of the figures correspond to loss, while the ones on top correspond to the κ values.

When visually examining the example outputs shown in Fig.4, we can confirm the quality of the numerical results. A slight oversegmentation is noticeable in all network outputs, making the vegetation appear thicker than in the ground truth. This, however, is not an obstacle to the deployment of the system in agricultural applications as the outputs will be processed in an object-based fashion. Classification imperfections can be noticed near the borders of the samples constructed from the input images. This can be remedied by presenting the classifiers with slightly overlapping samples if working at full image resolution, and is avoided completely when the classifier is trained to work with downsampled images (Lottes et al., 2018; Milioto et al., 2018). The output images from the *CA17* dataset confirm that the classifier copes well with scenes under substantial weed pressure, and is able to recognize the crop even when it is a part of a mixed dense vegetation patch.

Fig. 5 contains the graphical plots illustrating the progression of the loss function of the networks during training as well as of their classification performance on the test set for all datasets and experiment sets that have been conducted. In the case of training from scratch (Fig. 5(a)), regardless of the specific dataset, the loss function requires tens of thousands of iterations to decrease significantly, while test performance increases to acceptable levels. Both however fluctuate visibly. This is not surprising for several reasons. First, the context of the problem dictates small datasets, hence rendering the networks prone to overfitting.

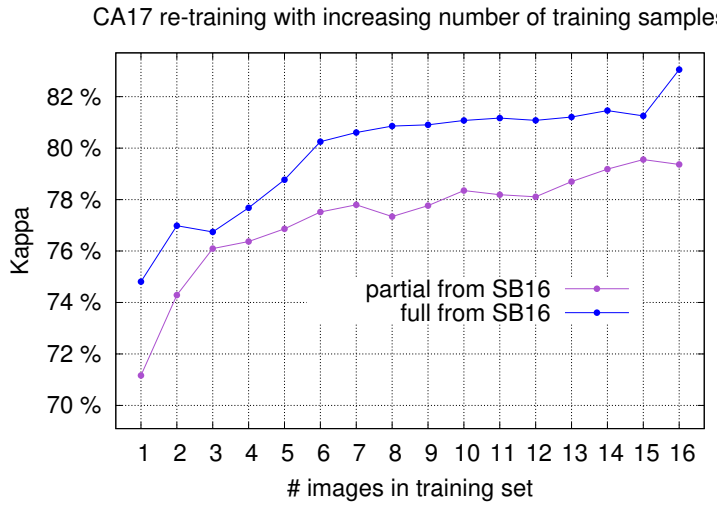

Figure 6: Performance of classifiers trained on *SB16*, and re-trained with *CA17-full* and *CA17-partial* for an increasing size of the training set.

Second, since our focus is on efficient transferability between crop types, the learning rate is relatively high. This was a deliberate decision based on our past experience (Atito et al., 2018; Ghazi et al., 2017), which we also confirmed empirically, as we aim to see how fast the network can reach practically acceptable levels of performance. And third, the datasets under study are truly challenging. Consequently, given the huge search space of the network it is not surprising that it occasionally switches between local optimal points and/or oscillates around them.

The retraining plots for the three datasets are shown in Fig. 5(b)–(d). The most significant observation is that with all three datasets, when training to recognise a crop using the features learned from another crop type, the convergence is faster by approximately an order of magnitude. This shows that the features a network learns for identifying a crop type from weeds can be effectively used in order to recognise a different crop type by focusing on adapting mostly its classifier weights to the new context. The specific amount of retraining effort to reach stable performance depends on the dataset, with *CA17* converging in the first few hundred iterations, while *ON17* and *SB16* reach training from scratch levels in a few thousand iterations. Furthermore, when dealing with the more challenging partial ground truth, the network, not surprisingly, has an increased difficulty in learning useful features, and is more prone to overfitting as the number of pixel labels per sample are even sparser. This is particularly visible for the *CA17-partial* data showing vegetation in later growth stages and thus more areas of mixed vegetation, and can be seen in Fig. 5(c).

In order to better asses the amount of data needed for re-training, we have additionally examined the performance of the classifier with smaller sizes of the training set. We have chosen the classifiers targeting *CA17*, our most challenging dataset, and observed the performance when the training set size varies from 1 to the 16 images used in the main experiments (where each *CA17* image corresponds to 16 samples). The performance in terms of the κ measure when the classifier is re-trained with both *CA17-full* and *CA17-partial* data is shown in Fig. 6. When retraining with *CA17-full* data, a rapid increase in performance can be observed up to 6 training images, and then again at 16 images. In case of retraining with *CA17-partial*, second and third training image cause the sharpest performance increase, however the performance continues steadily increasing with increasing the size of the training set. Both of these cases indicate that due to the variability in the *CA17* dataset, the classifier training could benefit from even more labelled data.

As further evidence that features learned by the network for one crop type are reused when retraining for a different target crop, in Fig. 7 we show the filters of the first convolution layer, typically responding to

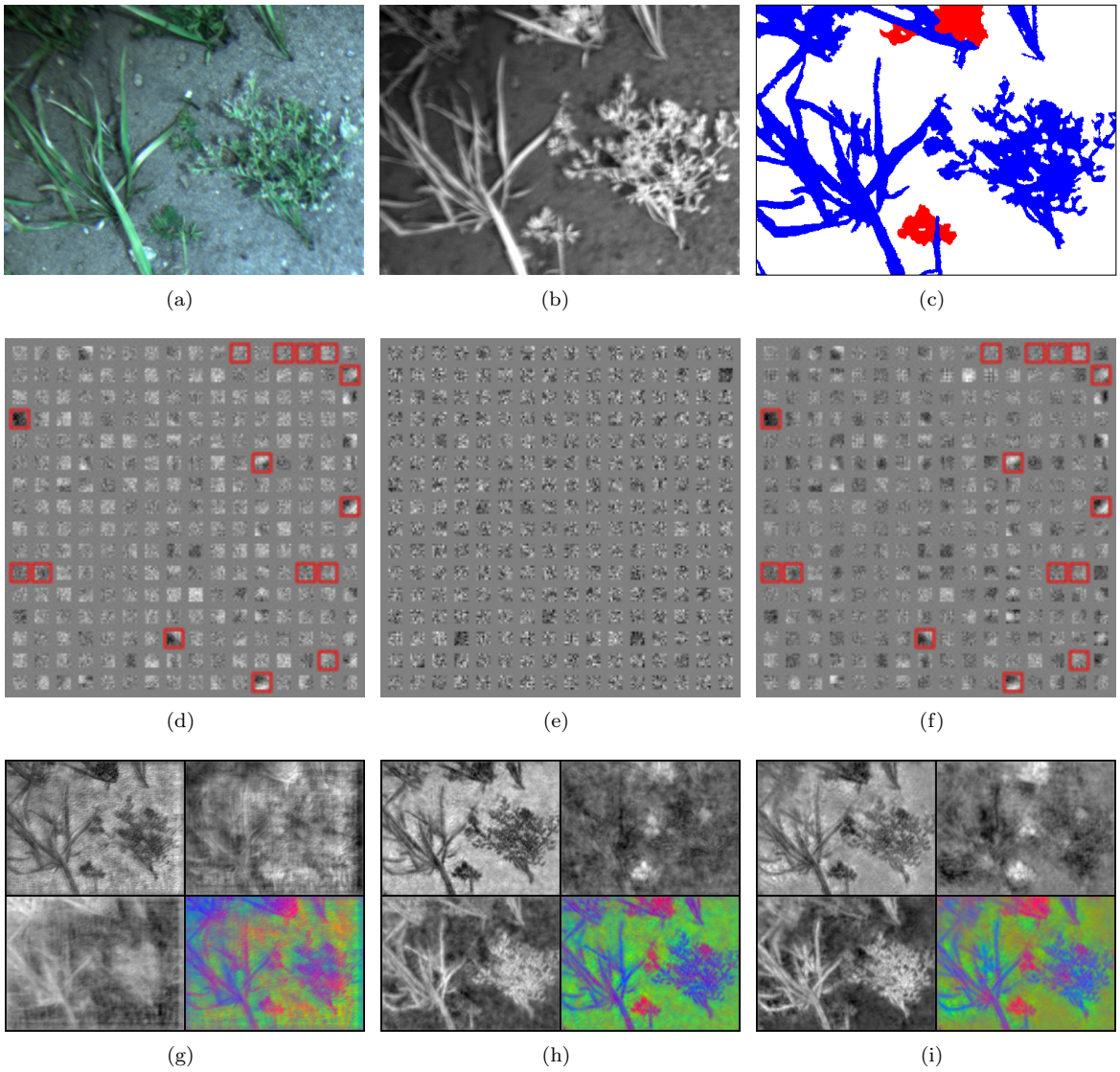

Figure 7: We show the filter weights and output probabilities per class for three different networks, when presented with images shown in (a) and (b). (d) and (g) correspond to a network trained from scratch on *SB16* dataset, (e) and (h) to a network trained from scratch on *CA17* dataset, and (f) and (i) to the network initialised from the *SB16* network, but retrained with *CA17-full* data. Selected similarities between filters in (g) and (i) are highlighted in red. The images in the last row show the predicted probability of the ground (top-left corner), crop (top-right) and weed classes (bottom-left), as well as these three probabilities overlayed into an RGB image (bottom-right), where the blue component corresponds to weeds and the red one to the target crop.

primitive image features, together with network outputs. We can observe that many of the filter weights of a network trained from scratch on *SB16* (Fig. 7(d)) are reused when the network is then retrained with *CA17-full* data (Fig. 7(f)), as a number of similar patterns can be noticed among the filter visualisation. At the same time, we can see that the retrained network does not learn features similar to a network trained from scratch on the same target crop, *CA17* (Fig. 7(e)). The visual appearance of the filter weights in Fig. 7(d) and 7(f), compared to those in Fig. 7(e), suggests that the network trained with the sugar beets crop learns features responding to different shapes, such as lines, corners and blobs (which are then used when the network is retrained for the carrot crop), while targeting the carrot crop during the initial network training results in the first convolution layer containing features corresponding to different kinds of textures. Despite relying on substantially different primitive features extracted by the first convolution layer, it can be seen from the network predictions that the outputs of the retrained network (cf. Fig. 7(i)) are much closer to the outputs of a network trained on the same target crop from scratch (cf. Fig. 7(h)) than they were before the retraining (as shown in Fig. 7(g)).

The classification of every 512×384 sized image lasted approximately 0.04 seconds, meaning that the processing of a single image from *SB16* required 0.16 seconds. We have thus achieved a modest processing frequency of 6.25Hz, while 10Hz has been typically considered to satisfy the real-time requirement on similar architectures (Milioto et al., 2018; Sa et al., 2018). This is due to our decision to use full-resolution images as we focused on transfer capabilities between different crop types, but can be improved significantly through first downsampling the original images (Lottes et al., 2018; Milioto et al., 2018).

5 Conclusions and Future Work

Given the recent success of methods based on deep convolutional neural networks for semantic segmentation of crops versus weed in precision agriculture, this article has addressed one of the main challenges preventing their wide scale deployment: their need for large quantities of labelled data. This poses a particular challenge in agriculture, as the environmental conditions can drastically change in a matter of hours, and similarly crop conditions can change within days, all the while data annotations need to be obtained rapidly after data acquisition.

To this end, we have investigated two original research directions. First, the possibility of fine-tuning, or more formally, of efficiently transferring the knowledge of a network trained on a given crop type, to another crop type. And second, whether this is possible with partially labelled images that can be prepared in relatively short and practically acceptable amounts of time.

To achieve our goal, we have employed a state of the art encoder-decoder CNN, and focused on three different crop types, using a dataset frequently used in the state of the art, and two that have been prepared by the authors and introduced for the first time to public use.

Our experiments have indicated that not only it is possible to retrain an already trained network with a new crop type, but that nearly optimal performance levels can be achieved with much less training effort as well. Moreover, another important conclusion has been that despite the expected decrease in performance when using partial ground truth, its extent is small enough to justify its practical interest, especially when considering the very significant gain in terms of on site data preparation duration.

Future work will focus on multiple directions: in the short-term, we plan first on developing efficient and effective partial ground truth preparation methods, suitable for on site use by non-experts through intelligent region annotation techniques. Secondly, after confirming transfer feasibility, we can now focus on improving the processing times to satisfy the real-time execution requirements and investigate the viability of transfer learning on downsampled images and between images of different spatial resolution. As a more long-term goal, we intend to establish a reliable solution framework that can accommodate and adapt to not only crop type changes but challenging real-world events as well, e.g. seasonal changes, diseases and pests.

References

Atito, S., Yanıkoglu, B., Aptoula, E., Ganiyusufoglu, I., Yildiz, A., Yildirir, K., and Baris, S. (2018). Plant identification with deep learning ensembles. *Working Notes of CLEF*, 2018.

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Bosilj, P., Duckett, T., and Cielniak, G. (2018a). Analysis of morphology-based features for classification of crop and weeds in precision agriculture. *IEEE Robotics and Automation Letters*, 3(4):2950–2956.

Bosilj, P., Duckett, T., and Cielniak, G. (2018b). Connected attribute morphology for unified vegetation segmentation and classification in precision agriculture. *Computers in Industry, Special Issue on Machine Vision for Outdoor Environments*, 98:226–240.

Brostow, G. J., Fauqueur, J., and Cipolla, R. (2009). Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2):88–97.

Chebrolu, N., Lottes, P., Schaefer, A., Winterhalter, W., Burgard, W., and Stachniss, C. (2017). Agricultural robot dataset for plant classification, localization and mapping on sugar beet fields. *Int. J. Robotics Research*.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Di Cicco, M., Potena, C., Grisetti, G., and 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)*, pages 5188–5195.

Eigen, D. and Fergus, R. (2015). Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *IEEE International Conference on Computer Vision*, pages 2650–2658.

Ghazi, M. M., Yanikoglu, B., and Aptoula, E. (2017). Plant identification using deep neural networks via optimization of transfer learning parameters. *Neurocomputing*, 235:228–235.

Guerrero, J., Pajares, G., Montalvo, M., Romeo, J., and Guijarro, M. (2012). Support vector machines for crop/weeds identification in maize fields. *Expert Systems with Applications*, 39(12):11149–11155.

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

Haug, S., Michaels, A., Biber, P., and Ostermann, J. (2014). Plant classification system for crop/weed discrimination without segmentation. In *Proc. WACV*, pages 1142–1149.

Hemming, J. and Rath, T. (2001). Computer-vision-based weed identification under field conditions using controlled lighting. *Journal of agricultural engineering research*, 78(3):233–243.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*.

Kusumam, K., Krajník, T., Pearson, S., Duckett, T., and Cielniak, G. (2017). 3d-vision based detection, localization, and sizing of broccoli heads in the field. *Journal of Field Robotics*, 34(8):1505–1518.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Lottes, P., Behley, J., Milioto, A., and Stachniss, C. (2018). Fully convolutional networks with sequential information for robust crop and weed detection in precision farming. *arXiv preprint arXiv:1806.03412*.

Lottes, P., Hörferlin, M., Sander, S., and Stachniss, C. (2017). Effective vision-based classification for separating sugar beets and weeds for precision farming. *J. Field Robotics*, 34(6):1160–1178.

McCool, C., Perez, T., and Upcroft, B. (2017). Mixtures of lightweight deep convolutional neural networks: applied to agricultural robotics. *IEEE Robotics and Automation Letters*, 2(3):1344–1351.

Milioto, A., Lottes, P., and Stachniss, C. (2018). Real-time semantic segmentation of crop and weed for precision agriculture robots leveraging background knowledge in CNNs. In *Proc. ICRA*.

Mortensen, A. K., Dyrmann, M., Karstoft, H., Jørgensen, R. N., and Gislum, R. (2016). Semantic segmentation of mixed crops using deep convolutional neural network. In *International Conference on Agricultural Engineering*.

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE transactions on Systems, Man, and Cybernetics*, 9(1):62–66.

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

Rouse Jr, J., Haas, R., Schell, J., and Deering, D. (1974). Monitoring vegetation systems in the Great Plains with ERTS. *NASA Special Publication-351*. presented at Third ERTS Symposium.

Ruiz-Ruiz, G., Gómez-Gil, J., and Navas-Gracia, L. (2009). Testing different color spaces based on hue for the environmentally adaptive segmentation algorithm (EASA). *Computers and Electronics in Agriculture*, 68(1):88–96.

Sa, I., Chen, Z., Popović, M., Khanna, R., Liebisch, F., Nieto, J., and Siegwart, R. (2018). weednet: Dense semantic weed classification using multispectral images and mav for smart farming. *IEEE Robotics and Automation Letters*, 3(1):588–595.

Slaughter, D., Giles, D., and Downey, D. (2008). Autonomous robotic weed control systems: A review. *Computers and electronics in agriculture*, 61(1):63–78.