Computers and Electronics in Agriculture xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Original papers

Synthetic bootstrapping of convolutional neural networks for semantic plant part segmentation

R. Barth[a,c,*], J. IJsselmuiden[b], J. Hemming[a], E.J. Van Henten[b]

[a] Wageningen University & Research, Greenhouse Horticulture, P.O. Box 644, 6700 AP Wageningen, The Netherlands
[b] Wageningen University & Research, Farm Technology Group, P.O. Box 16, 6700 AA Wageningen, The Netherlands c Harvard University, Biorobotics Laboratory, 60 Oxford street, Cambridge, MA, United States of America.
[c] Harvard University, Biorobotics Laboratory, 60 Oxford street, Cambridge, MA, USA

ARTICLE INFO

Keywords:
Computer vision
Semantic segmentation
Synthetic dataset
Bootstrapping
Big data

ABSTRACT

A current bottleneck of state-of-the-art machine learning methods for image segmentation in agriculture, e.g. convolutional neural networks (CNNs), is the requirement of large manually annotated datasets on a per-pixel level. In this paper, we investigated how related synthetic images can be used to bootstrap CNNs for successful learning as compared to other learning strategies. We hypothesise that a small manually annotated empirical dataset is sufficient for fine-tuning a synthetically bootstrapped CNN. Furthermore we investigated (i) multiple deep learning architectures, (ii) the correlation between synthetic and empirical dataset size on part segmentation performance, (iii) the effect of post-processing using conditional random fields (CRF) and (iv) the generalisation performance on other related datasets. For this we have performed 7 experiments using the *Capsicum annuum* (bell or sweet pepper) dataset containing 50 empirical and 10,500 synthetic images with 7 pixel-level annotated part classes. Results confirmed our hypothesis that only 30 empirical images were required to obtain the highest performance on all 7 classes (mean IOU = 0.40) when a CNN was bootstrapped on related synthetic data. Furthermore we found optimal empirical performance when a VGG-16 network was modified to include à *trous* spatial pyramid pooling. Adding CRF only improved performance on the synthetic data. Training binary classifiers did not improve results. We have found a positive correlation between dataset size and performance. For the synthetic dataset, learning stabilises around 3000 images. Generalisation to other related datasets proved possible.

1. Introduction

1.1. Research aim

In this paper we investigated a methodology to reduce the dependancy on manually annotated datasets for plant part segmentation in agriculture when applying state-of-the-art deep learning methods, e.g. convolutional neural networks (CNN) for semantic segmentation. CNNs were bootstrapped by a synthetic dataset and fine-tuned on a small manually annotated dataset. Additionally, our aim was to further specify CNN and data requirements for this task and therefore we investigated (i) the correlation between synthetic dataset size and performance, (ii) the minimum required amount of fine-tuning data, (iii) explicit improvements for this task by training part classes separately in binary classifiers, (iv) the effect of post-processing using conditional random fields (CRFs) and (v) the generalisation power to related datasets differing in acquisition distance and hardware.

Currently state-of-the-art computer vision methods for semantic segmentation are dominated by supervised machine learning such as CNNs (Everingham et al., 2015; Zhao et al., 2016; Wu et al., 2016). With the advent of these methods comes the requirement of large and detailed annotated datasets (Najafabadi et al., 2015). Although this depends on the model's number of free parameters and the problem complexity, it has already been shown that dataset size and classification performance are positively correlated (Banko and Brill, 2001; Brants et al., 2007). Specifically for deep learning methods, this correlation holds given sufficient model size, training iterations and regularisation (Erhan et al., 2010; Srivastava et al., 2014).

Unfortunately, the lack of annotations frequently imposes a new bottleneck for learning. Annotating per pixel class labels is labour intensive, which can become infeasible for large sets with a multitude of classes. Particularly computer vision in domains like agriculture, with a high amount of occlusions and high object and environmental complexity (Gongal et al., 2015), obtaining detailed annotated datasets that

* Corresponding author at: Wageningen University & Research, Greenhouse Horticulture, P.O. Box 644, 6700 AP Wageningen, The Netherlands.
E-mail addresses: ruud.barth@wur.nl (R. Barth), joris.ijsselmuiden@wur.nl (J. IJsselmuiden), jochen.hemming@wur.nl (J. Hemming), eldert.vanhenten@wur.nl (E.J. Van Henten).

https://doi.org/10.1016/j.compag.2017.11.040
Received 14 July 2017; Received in revised form 23 October 2017; Accepted 30 November 2017

capture all image variance often proved to be insurmountable. One solution is to perform large-scale crowd-sourcing (Kavasidis et al., 2014; Everingham et al., 2010a; Russell et al., 2008), though this can remain costly. Another popular and successful solution is to bootstrap (or pre-train) machine learning models with synthetic data (Ros et al., 2016; Shotton et al., 2013; Cordier et al., 2016).

To investigate this solution and its boundary conditions in the domain agricultural computer vision, we report on deep learning experiments that used the *Capsicum annuum* (bell or sweet pepper) dataset (Barth et al., 2017) containing 8 classes (7 plant parts plus background class) annotated both in empirical images (50) and corresponding synthetic images (10,500). The dataset provides a demarcated scope for semantic part segmentation, comparable to datasets like The Penn-Fudan face part dataset (7 classes) (Wang et al., 2007), Labeled Faces in the Wild (3 classes) (Learned-Miller et al., 2016), Caltech-UCSD Birds-200-2011 bird part dataset (3 classes) (Wah et al., 2011) and the Pascal-Person-Part dataset (6 classes) (Chen et al., 2014, 2016a).

The need for object and part recognition on a per-pixel level in this domain arises from requirements in harvest robotics (Bac et al., 2013), phenotyping (van der Heijden et al., 2012) and disease detection (Polder et al., 2014), which require precise object classification and localisation. For example in harvest robotics, obstacle maps for motion planning need to have a resolution up to the plant part level (Bac et al., 2014a, 2016).

1.2. Hypotheses

Our main hypothesis was that only a small manually annotated empirical fine-tuning dataset would remain required for optimal and successful empirical learning after bootstrapping a convolutional neural network (CNN) with related synthetic data, as compared to other training methods.

To gain further insights in the dataset size requirements for learning, our additional hypothesis was that segmentation performance increases with larger synthetic or empirical dataset size, though will level out at a certain order of magnitude.

During the experiments we developed further additional hypotheses and tests for causes of certain segmentation results, e.g. the training of binary classifiers to circumvent skewed class learning and the addition of CRFs to improve local class segmentation boundaries.

Finally, we hypothesise that a empirically fine-tuned model can robustly generalise to related datasets, due to the distributed and hierarchical representation learning of CNNs. We define related datasets as images of the same crop, though under different conditions such as illumination, acquisition hardware or imaging distance.

Although previously we have provided brief evidence for our main hypothesis as a perspective towards future research (Barth et al., 2017), in this paper we will expand on that work. We affirm our earlier experiments with a more advanced CNN architecture and place the results in a broader context in this paper.

1.3. Requirements

Regarding our main hypothesis, we require a quantification of the small empirical dataset size. Although this number is arbitrary, we aim for a supervised machine learning methodology that needs no more of a manual annotation effort than 2 days. Given an average previously reported annotation time of 30 min per image in the used dataset (Barth et al., 2017), this translates to an upper bound of 30 images.

For optimal learning, we require that no other learning scheme that includes combinations of synthetic, empirical, other related and/or unrelated data for bootstrapping and/or fine-tuning has a higher performance on the empirical test set.

We define successful learning as the recognition of all classes, preferably with a low performance variance amongst classes. Success itself is quantified using the intersection-over-union measure, stated in Eq. 1. However, because the extent of success is highly task dependent, we do not define hard requirement values. We can however indicate that for tasks such as detection, a relatively low ($IOU \geqslant 0.5$) is sufficient because it is not the precise overlap that counts, but partial recognition suffices. However, for tasks such as phenotyping the measure is required to be high ($IOU \geqslant 0.9$), since exact dimensions and morphology are of interest.

1.4. Contributions

Our work provides the field of computer vision in agriculture a pioneering methodology for state-of-the-art segmentation, whilst simultaneously reducing the constraints on labour intensive manual annotations. Results are a key part in the next leap of robotization in agriculture to keep up with the increasing demand of productivity and quality whilst decreasing the pressure on resources required (Bac et al., 2014b).

1.5. Research context

The use of synthetic image data is emerging as a powerful tool in the computer vision community to generate training data for bootstrapping machine learning models. Such models can either be co-trained or fine-tuned with empirical data (Dittrich et al., 2014; Kondaveeti, 2016), on which the models are also deployed. Examples that show improved object recognition performance can be found in multiple domains, e.g. 3D human pose estimation from depth images (Shotton et al., 2013) and multi-modal magnetic resonance imaging for pathological cases (Cordier et al., 2016). Other notable examples for urban scene classification showed that synthetic images alone were sufficient as training data for a model applied to real scenes (Hattori et al., 2015), though accuracy was increased when combined with real training data (Ros et al., 2016).

To a certain extent, synthetic training and empirical fine-tuning can be seen as a form of soft transfer learning (Caruana, 1995; Bengio, 2011; Bengio et al., 2011), where a model can be successfully applied to a different task and domain. However, since in our case not only the task is equal but also the data is highly similar, we therefore adhere to the term bootstrapping.

The challenge of semantic segmentation in computer vision can be described as to dividing an image into non-overlapping meaningful regions, ultimately determining the object (part) class on a per-pixel level. Historically, semantic segmentation has been performed mainly supervised, although weakly or non-supervised approaches have also been successful for certain problems (Wehrens, 2010; Zhu et al., 2016). Compared to other methods that only generate a single high-level label description per image, semantic segmentation has the benefit of localisation in the image plane. This advantage is useful for applications such as robotics where object manipulation and navigation is dependent on accurate positional information, e.g. autonomously driving cars (Shapiro, 2016; Badrinarayanan et al., 2017), warehouse order picking robots (Zeng et al., 2016) or agricultural robots (Bac et al., 2013, 2016).

Specifically for agricultural robotics, per-pixel level annotations are either required or will improve performance as opposed to using coarse localisation methods (e.g. bounding box detection). We see the main applications in (i) crop handling, (ii) phenotyping and (iii) disease detection.

Regarding crop handling, harvesting robotics requires precise end-effector placement at the target fruit (Bac et al., 2016; Li et al., 2016) or near the fruit such as a peduncle (Sa et al., 2017; Henten et al., 2009) or to avoid obstacles during that motion (Bac et al., 2013). Per-pixel class segmentations can provide the information to allow this precise end-effector placement when registered depth can be inferred by using 3D sensors, e.g. stereo imaging (Bac et al., 2014a) or time-of-flight cameras (van der Heijden et al., 2012). Other crop tasks, such as open field weeding, also require to differentiate on a per-pixel level where the

crop and weeds are (Milioto et al., 2017a,b) to allow for precise spraying (de Soto et al., 2016) or dutch hoeing (Hemming and Rath, 2001). Also for the task of leaf picking, per-pixel labels might be beneficial regarding precision over current bounding box approaches (Ahlin et al., 2016).

Regarding automated phenotyping, where the task is to correlate plant parameters with their underlying genetics to guide plant breeding, the localisation of plant parts in the image is a hard requirement (Araus and Cairns, 2014). Plant parameters such as leaf size (van der Heijden et al., 2012), stalk thickness (Vijayarangan et al., 2017) or spikelet counts (Pound et al., 2017) are to be estimated with high precision. Per-pixel segmentation is a key development for this domain. However, also for this task a registration with depth information is required to infer real world dimensions from the image coordinates that the segmentation provides.

For plant disease detection, the task is not only to determine which plant is healthy or diseased but moreover where on the plant the infection is present (Phadikar and Sil, 2008) to allow local automated treatment (Oberti et al., 2013) or to map the phase or size of the infection in the crop to guide crop management (Lu et al., 2017).

Previously successful semantic segmentation methods were based on manually crafted features as input for shallow learning models such as support vector machines or random forests (Johnson et al., 2013; Fulkerson et al., 2009). For other computer vision tasks, similar methods were recently superseded by convolutional neural networks (Everingham et al., 2015). However, initially such methods could not perform semantic segmentation due to the convergent nature of the networks' architectures. Convolutional neural networks start with global information that is compressed in an increasingly spatially independent hierarchy of features, forming a distributed representation of the input whilst losing locality information. However, for semantic segmentation preserving the locality information is key.

At first, solutions for the locality problem appeared by learning additional objectives like coarse bounding boxes that represented the location of the object in the image (Uijlings et al., 2013; Pont-Tuset et al., 2015). Although often still preferred for speed or annotation costs, the downside of bounding box methods is the lack of segmentation detail. Later approaches tried merging classifications from multiple levels in the network's hierarchy combined with super-pixel pre-segmentation (Farabet et al., 2013). Recently, a novel architecture was presented using fully convolutional neural networks (Papandreou et al., 2015; Chen et al., 2015), differing from other networks by replacing the fully connected layers with convolutional ones and adding dense predictions using the à *trous* algorithm (meaning with holes, also reported as atrous) (Mallat, 1999). Furthermore, not uncommon with previous approaches, an integrated layer was added for applying a CRF as post-processing to refine the lost locality of the segmentation.

To further expand the challenge of object localisation, efforts were made to localise parts within those objects (Felzenszwalb and Huttenlocher, 2000, 2005), later also contrived for semantic segmentation (Wang and Yuille, 2015; Tsogkas et al., 2015). Although this refinement can be considered as merely increasing the number of classes, this would neglect the strong spatial correlations between object parts. Some methods applied compositional models and high-level information to include these relationships and improve on object part segmentations. With convolutional neural networks however, the distributed hierarchical representation of objects and their features facilitates learning these correlations.

This paper will first describe the research materials in Section 2. The general methodology across experiments is then described in Section 3. In the sections that follow, each experiment is reported separately with their own introduction, method, results and discussion section. A general discussion is provided in Section 11 and we conclude the paper in Section 12.

Fig. 1. Examples of empirical (top) and synthetic (bottom) color images (left) and their corresponding ground truth labels (right). Class labels: ● background, ● leafs, ● peppers, ● peduncles, ● stems, ● shoots and leaf stems, ○ wires and ● cuts where pepper where harvested. (For interpretation of the reference to colour, the reader is referred to the web version of this article.)

2. Materials

2.1. Dataset description

The *Capsicum annuum*, sweet- or bell pepper image dataset (Barth et al., 2017) consists of 50 empirical images of a crop in a commercial high-tech greenhouse and 10,500 corresponding synthetic images, modelled to approximate the empirical set visually and geometrically. The synthetic images were generated to reflect the empirical situation by rendering random 3D meshes of plants. These meshes were randomly generated using 21 empirically measured plant parameters. To create realistic images, the greenhouse growing architecture was modelled as well as similar camera and illumination settings for rendering.

In both image sets, 8 classes were annotated on a per-pixel level, either manually for the empirical dataset or automatically for the synthetic dataset. In Fig. 1 examples of images of the dataset are shown. The dataset was publicly released at: http://dx.doi.org/10.4121/uuid:884958f5-b868-46e1-b3d8-a0b5d91b02c0.

2.2. Convolutional neural network architectures

For our experiments we used the publicly available fully convolutional neural network (CNN) architectures of DeepLab (Papandreou et al., 2015; Chen et al., 2015) implemented on top of Caffe (Jia et al., 2014). Although other deep learning implementations were being researched for semantic segmentation (Shelhamer et al., 2016; Mostajabi et al., 2014; Long et al., 2015), those were not yet available for verification at the time of our research. DeepLab models can either be trained with weakly semi-supervised learning (e.g. bounding boxes) or with strong supervision (e.g. per-pixel labels). For our approach, the detailed annotated dataset allowed training strong supervision models, resulting in more localised labeling as compared to bounding boxes.

The underlying architecture for the CNN models was based on VGG-16 (Simonyan and Zisserman, 2014) as depicted in Fig. 2. VGG-16 was originally intended for global object detection. To adjust it for semantic segmentation, 2 changes were made to the architecture (Papandreou et al., 2015). First the fully connected multi-layered perceptron at the end of the network was replaced by fully convolutional layers (Long et al., 2015); hence the network was applied in a convolutional manner on the input image at its original resolution which resulted in per-pixel labels. However, given the used stride of the convolutions this resulted in down-sampled prediction. Therefore the second adjustment was

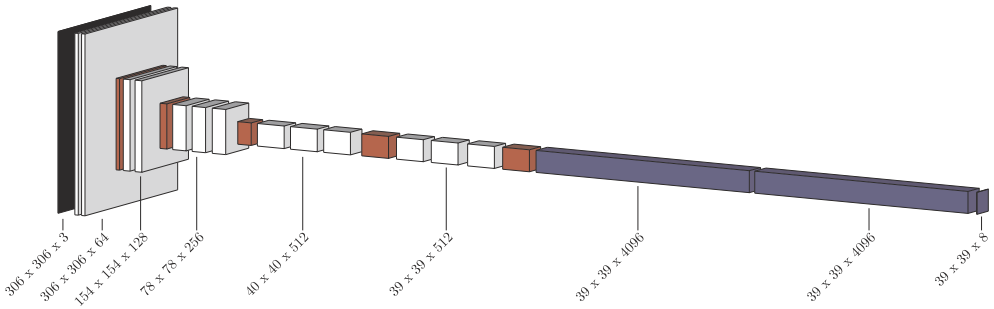

Fig. 2. Convolutional neural network architecture of Deeplab Vanilla model, based on the VGG-16 composition. All models used in this paper are modified versions of this architecture. Color encodes layer type: ● cropped input image, ○ convolutional layer + rectified linear unit, ● pool layer and ● fully connected layers transformed to fully convolutional ones. Dimensions indicate feature map width × height × number of feature maps. (For interpretation of the references to colour, the reader is referred to the web version of this article.)

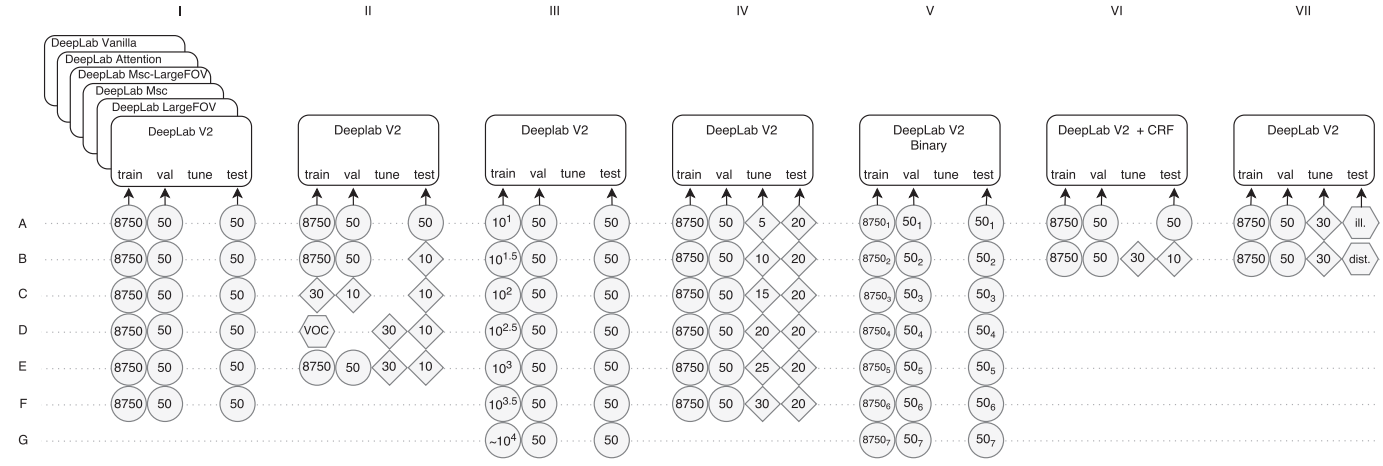

Fig. 3. Overview of performed main experiments I through VII, with sub-experiments A through G. Model architectures (rectangles) were trained, validated, tuned or tested with dataset types empirical (diamond), synthetic (circle) or different but related datasets (hexagon) with the number of image samples displayed within. A subscript indicates that only a single specific class was used.

made by implementing the à *trous* algorithm; by upscaling the filters with filler zeros, predictions at the original resolution could be made (Chen et al., 2016a). Although potentially a better performing RESNET-101 implementation for DeepLab also exists (Chen et al., 2016a), we were not able to train such network due to GPU memory constraints.

To the base model DeepLab-Vanilla (Papandreou et al., 2015), the following additional adjustments to the architecture were previously made to explore the effect on segmentation performance on the PASCAL VOC 2012 dataset (Everingham et al., 2010b).

1. **Increasing the field-of-view (DeepLab-LargeFOV).** Adjusted input stride to 12 and using a kernel size of 3 × 3 at the first fully convolutional layer, the receptive field was doubled in comparison to DeepLab-Vanilla whilst having a third of the number of free parameters (Chen et al., 2015).
2. **Addition of multi-scale predictions (DeepLab-MSc)**. Improved segmentation accuracy at the boundaries of objects, the final feature map layer receives 5 additional feature maps convoluted from intermediate layers in the network (Chen et al., 2015).
3. **Combination of the former two adjustments (DeepLab-MSc-LargeFOV)** (Chen et al., 2015).
4. **Addition of an attention model on multiple scales (DeepLab-Attention).** The input image was resized to several scales and used for training both a network and an attention model. The attention model weighed each image scale and each feature thereof for the final segmentation (Chen et al., 2016b).
5. **Addition of *à trous* spatial pyramid pooling (ASPP) (Deeplab-v2).** ASPP included image context at multiple scales by convolutional feature layers with different fields-of-view (Chen et al., 2016a; He et al., 2014).

For each model, pixels were independently classified without considering label agreement across the image. Applying a fully connected CRF (Krähenbühl and Koltun, 2011) can include long-range object dependencies and reduces label noise while refining part edge details. The CRF takes the CNN pixel prediction as the unary potentials and maximises label consistency by encouraging the assignment of locally similar labels that have similar properties. The DeepLab models can be extended with such a CRF as an extra integrated layer, although its parameters cannot be trained by back-propagation and should be found separately.

2.3. Hardware

Experiments were run on a NVIDIA DevBox system with 4 TITAN X Maxwell 12 GB GPUs, Intel Core i7-5930 K and 128 GB DDR4 RAM running Ubuntu 14.04. As a dependancy for the DeepLab V2 Caffe version, the archived version of CUDA 7.5 was installed. Training a single model took 24 h on average (using a batch size of 10 cropped images of 300 × 300 pixels per GPU and 30,000 iterations). Testing on a single 800 × 600 pixel image took around 200 ms.

3. Methods

We performed Experiments I through VII, each consisting of sub-experiments by varying dataset composition and/or model architecture. In Fig. 3 an overview of the main experiments is presented. Not all permutations of models, hyper-parameters, dataset types and sizes were explored due to the infeasible combinatorial computational cost. For this reason, the best performing model architecture was selected first by evaluating Experiment I, which was then further used for Experiments II-VII.

For each experiment, the same range of images was selected for the train, validation or test phase to make sure unique synthetic plant

models or empirical plants were separated in the different phases and equal between experiments. The synthetic dataset consisted of 6 scenes of 1750 images (10,500 total), with each scene containing unique plants (Barth et al., 2017). To ensure separation between unique images, the first 5 scenes with images 1–8750 were used as synthetic training images whereas the remaining scene were used for validation images (8851–8900) and test images (8751–8800). Similarly, for the empirical dataset, images 1–30 of unique plants were used for training and the remainder images 31–40 for validation and 41–50 for testing.

The hyperparameters of the network were manually optimised using the validation dataset and a combination of models and dataset configurations as suggested by (Goodfellow et al., 2016; Bengio, 2012). Specifically, we searched for a learning rate that reduced the loss on the validation set gradually over the iterations towards zero. For the solvers Stochastic Gradient Decent, AdaDelta, Adaptive Gradient Descent, ADAM, Nesterov, and RMSprop, learning rates of 0.1, 0.01, 0.001 and 0.0001 were explored. We noted that none of the models overfitted on the validation set, though IOU performance differed between solvers and learning rates.

The hyperparameter search resulted in the choice of Adaptive Moment Estimation (ADAM) (Kingma and Ba, 2014) with $\beta_1 = 0.9, \beta_2 = 0.999, \varepsilon = 10^{-8}$ and a base learning rate of 0.001 for 30,000 iterations with a batch size of 4. These chosen hyper-parameters were found to be consistently optimal for multiple experiments with different datasets and therefore we fixed them across all experiments. To each model, an adjustment was made in the layer weight initialisation procedure. We updated the models to using MSRA weight fillers (He et al., 2015; Mishkin and Matas, 2015). Furthermore, the dropout rate (Srivastava et al., 2014) was adjusted to 0.50 to improve generalisation.

As a performance measure the intersection-over-union (IOU) was used, as described in (He and Garcia, 2009; Everingham et al., 2010a; Barth et al., 2017) which is also known as the Jaccard Index similarity coefficient. The IOU can be determined per class or as an average over all classes. The measure as an average over all classes is defined in Eq. (1), where for each class their IOU equals the intersection of the semantic segmentation and the true labels divided by their union. To derive the measure, first a pixel-level confusion matrix C is calculated first for each image I in dataset D, where $S^I_{gt}(p)$ is the ground truth label of pixel p in image I and $S^I_{ps}(p)$ is the predicted label. This implies that C_{ij} equals the count of pixels with ground truth label i and prediction j.

$$IOU = \frac{1}{L}\sum_{i=1}^{L}\frac{C_{ii}}{G_i + P_i - C_{ii}}, \quad \text{where} \tag{1}$$

$$C_{ij} = \sum_{I \in D} |\{p \in I | S^I_{gt}(p) = i \wedge S^I_{ps}(p) = j\}|, and \tag{2}$$

$$G_i = \sum_{j=1}^{L} C_{ij} \quad \text{and} \quad P_j = \sum_{i=1}^{L} C_{ij} \tag{3}$$

Hence G_i denotes the total number of pixels labeled with class i in the ground truth and P_j the total number of pixels with prediction j in the image.

Apart from quantitative evaluation, qualitative evaluation of the segmented images was performed to assess differences in segmentation style (e.g. course versus fine). Albeit two different models can achieve equal IOUs, the underlying distributions can be distinctive. Furthermore, the emergent classification property of the spatial distribution of the true and false positives can be determined this way.

Aside from previously mentioned regularisation methods, overfitting for each experiment was prevented by selecting the optimal model by periodically checking the performance on the validation set. This method of early stopping requires to select a point where performance either stabilised or decreased. This was done manually by evaluating the IOU per class over the iterations. In this research, the early stopping point found for each model was at 30.000 iterations.

4. Experiment I

CNN model architecture is a key factor to classification performance (Bengio, 2009). Proven base architectures are often modified with new insights to validate the enhancements on a range of benchmark datasets. To investigate how a set of model architecture modifications relate to part segmentation performance for our use-case, we compared 6 deep learning architectures in Experiments I-A through I-F, ordered by increasing expected performance according to previously obtained results by their authors. From these experiments we selected the best performing model for further Experiments II-VII. Assuming that the largest dataset for training results in optimal performance on the test set (Soekhoe et al., 2016), (a hypothesis we aimed to verify for our case with Experiment III) the full synthetic training dataset of 8,750 images was used to train each model in Experiment I. We did not use empirical data for Experiment I since we assumed performance and behaviour of the different architectures would be ranked similarly given comparable domain images, close on the same manifold.

4.1. Methods

The following experiments were run using synthetic images 1–8750 for training, validation images 8851–8900 and testing images 8751–8800; I-A: DeepLab-Vanilla, I-B: DeepLab-MSc, I-C: DeepLab-LargeFOV, I-D: DeepLab-MSc-LargeFOV, I-E: DeepLab-Attention and I-F: DeepLab-V2.

The performance was compared quantitatively by evaluating mean IOUs over the test set images and over each class. Furthermore, the final segmentations were assessed qualitatively. The primary requirement for a model to be selected for future experiments was the ability to recognise all classes. The secondary requirement was high IOU performance relative to the other models.

4.2. Results

In Fig. 4 the performances of model I-A through I-F are shown. For qualitative investigation, an example segmentation for each model with corresponding color and ground truth image is shown in Fig. 5. Furthermore, the underlying per class probability heat maps are presented to provide insight into the raw output of the CNN.

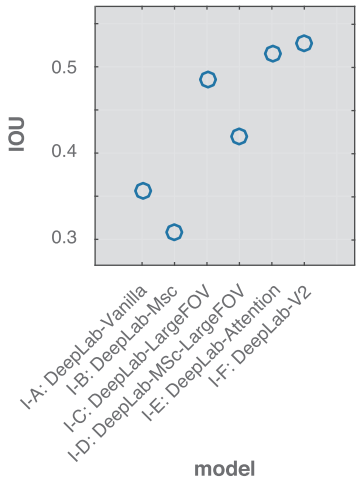

Fig. 4. Results of Experiment I, displaying mean test set IOU over each class for each model architecture I-A through I-F.

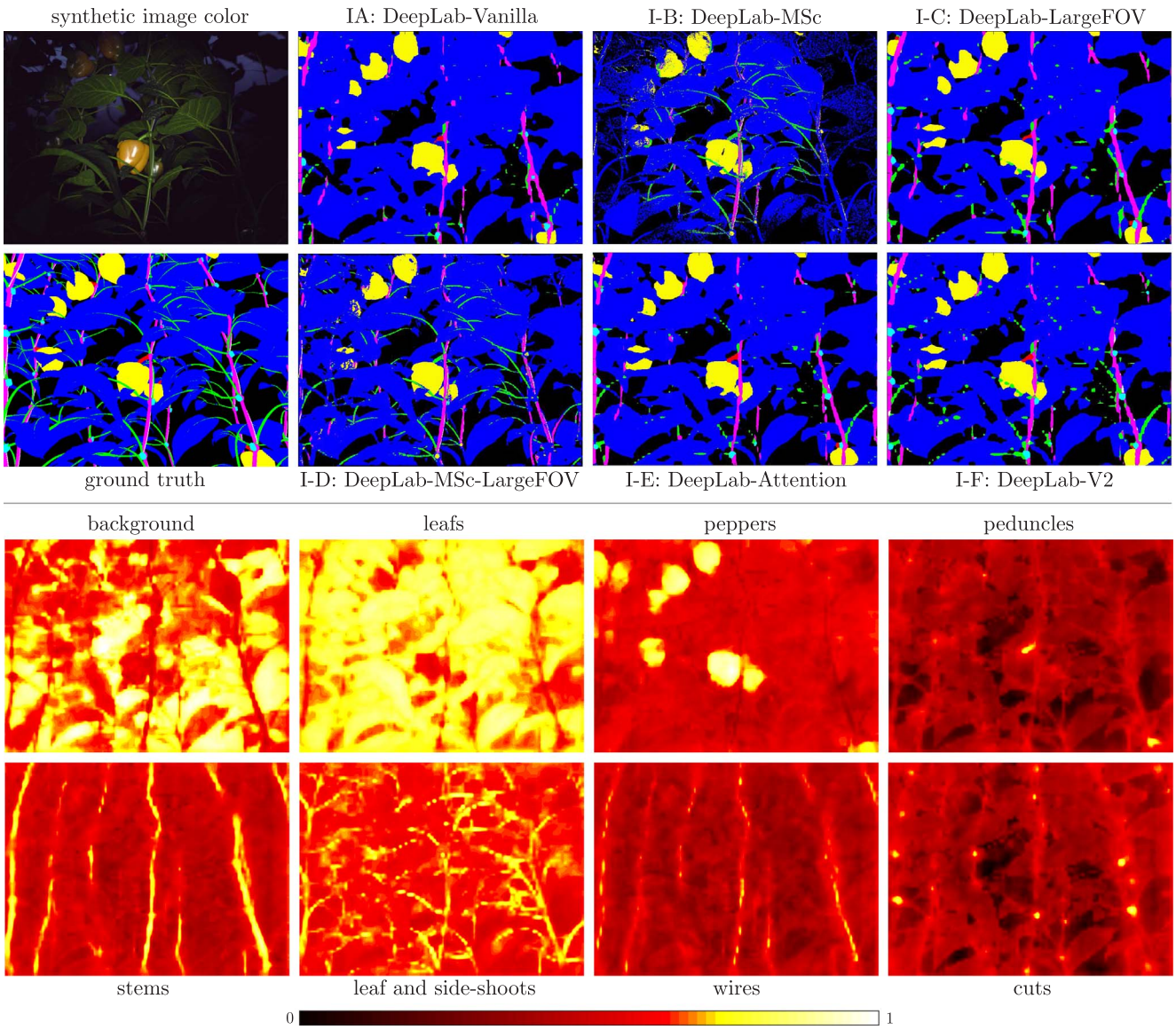

Fig. 5. Results of Experiment I. The top two rows show segmentation results of Experiments I-A through I-F. Class labels: ● background, ● leafs, ● peppers, ● peduncles, ● stems, ● shoots and leaf stems, ○ wires and ● cuts. The bottom two rows show per class probability maps of DeepLab-V2 for the synthetic color image from which the final segmentation was derived by selecting per pixel the class with the highest value of all maps. (For interpretation of the references to colour, the reader is referred to the web version of this article.)

4.3. Discussion

In Fig. 4 we observed that all of the implemented modifications to the base model (DeepLab-Vanilla) improved the mean IOU results, except for adding 5 feature maps convoluted from intermediate layers in the final feature map (MSc) (Chen et al., 2015). Evaluating Fig. 5, the models fail to learn the uncommon classes or darker areas in the image, e.g. wires and cuts. The commonality in the datasets was determined as the percentage of ground truth pixels (in decreasing order): background, leafs, peppers, shoots and leaf stems, main stems, wires, peduncles and cuts. This is further supported with results provided in Appendix A), where the IOU per class on the validation set can be observed over the training time for each model.

Fig. 5 shows the qualitative performance differences between models. DeepLab-MSc had the desired property of sharp segmentation boundaries, though failed to cope with pixels lacking distinct color in the outer area of the image. Furthermore, combining MSc with LargeFOV resulted in the neglect of uncommon classes, as can be observed in I–D in Appendix A. The other models appeared coarse around the plant parts, providing more true positive detections.

In Fig. 5 the per class probability maps by DeepLab-V2 for a synthetic example image is shown. These distributions gave insight in the underlying learned class probabilities. It shows that the stem and wire classes were highly overlapping and the final segmentation of the wires was often overruled by the stem class. Furthermore, although leaf stems and side-shoot segmentations were sparsely present in the final segmentation, the model appeared to learn the individual probability distributions quite well. Plausibly, learning binary classifiers for these classes should improve IOU performance, which we investigated in Experiment V.

Overall, the most recent proposed modification DeepLab-V2; the addition of à *trous* spatial pyramid pooling (Chen et al., 2016a), had the highest IOU. Moreover, it is being able to learn all plant parts (see Appendix Appendix A) hence meeting one of our requirements. We therefore selected the DeepLab-V2 model for Experiments II-VII.

Future research is suggested of adding the beneficial DeepLab-MSc sharp edge properties to DeepLab-V2, without suppressing the uncommon classes.

5. Experiment II

We hypothesise that synthetic bootstrapping and fine-tuning with a small empirical dataset can improve performance over other learning strategies. Previously we explored this briefly using the DeepLab-Vanilla model (Barth et al., 2017). In this paper we try to further validate those results and expand on that work. We ran the following 5 experiments, using the DeepLab-V2 model.

5.1. Methods

The motivation for each experiment is given below and the used image indices are shown between brackets. To evaluate the performance, the mean IOU over the test set images and over all classes for each experiment was obtained.

II-A DeepLab-V2. *Train: synthetic (1–8750). Test: synthetic (8751–8800).*
This experiment was run to obtain a baseline performance of the model when having access to a large and detailed annotated dataset for this domain. Assuming performance increased with dataset size until the model's complexity was saturated (Zeiler and Fergus, 2014), this experiment provides insight into the theoretical upper bound of the performance of all experiments in II. This assumption was further tested in Experiment III.

II-B DeepLab-V2. *Train: synthetic (1–8750). Test: empirical (41–50).*
Determines to what extent a synthetically trained model can generalise to a similar set in the same domain (e.g. empirical images) without fine-tuning. If performance would approximate the performance obtained in I-F, this is evidence against the aspect of our main hypothesis that fine-tuning improves performance.

II-C DeepLab-V2. *Train: empirical (1–30). Test: empirical (41–50).*
Investigates if the model can learn with only a small empirical dataset. If performance would approximate results of Experiment I-F, this is evidence against the aspect of our main hypothesis that a *large* dataset for bootstrapping would be required for improved performance.

II-D DeepLab-V2. *Train: PASCAL VOC. Fine-tune: empirical (1–30). Test: empirical (41–50).*
Compares the effect of bootstrapping with a non-related dataset. If the performance approximate the performance obtained in I-F, this is evidence against the aspect of our main hypothesis that a *related* dataset of the same domain is needed for improved performance.

II-E DeepLab-V2. *Train: synthetic (1–8750). Fine-tune: empirical (1–30). Test: empirical (41–50).*
Assesses the performance of bootstrapping with a related dataset and fine-tuning with a small empirical set. Given our main hypothesis, this experiment is expected to achieve best performance on empirical data.

5.2. Results

The IOU results for each experiment are shown in Fig. 6 and were split into per class IOUs in Fig. 8. Segmentation results for the best performing model on synthetic data II-A and empirical data II-E are shown in Fig. 7.

5.3. Discussion

From the quantitative results in Fig. 6 we derived the following:

II-A This model indicated a benchmark or baseline of optimal performance when the model had access to a large dataset with perfect ground truth. We assumed a positive correlation between dataset size and classification performance (Banko and Brill, 2001; Brants et al., 2007), to be further validated in Experiment III.

II-B Without fine-tuning, the synthetically bootstrapped model could not generalise properly to empirical data.

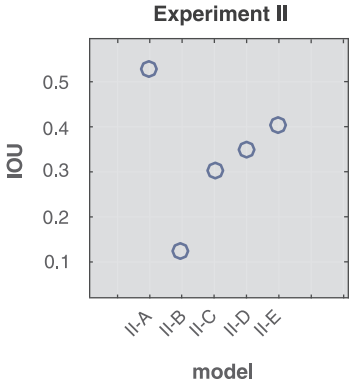

Fig. 6. Results of Experiment II-A through II-E. The mean IOU over the test set images and over all classes for each experiment is displayed.

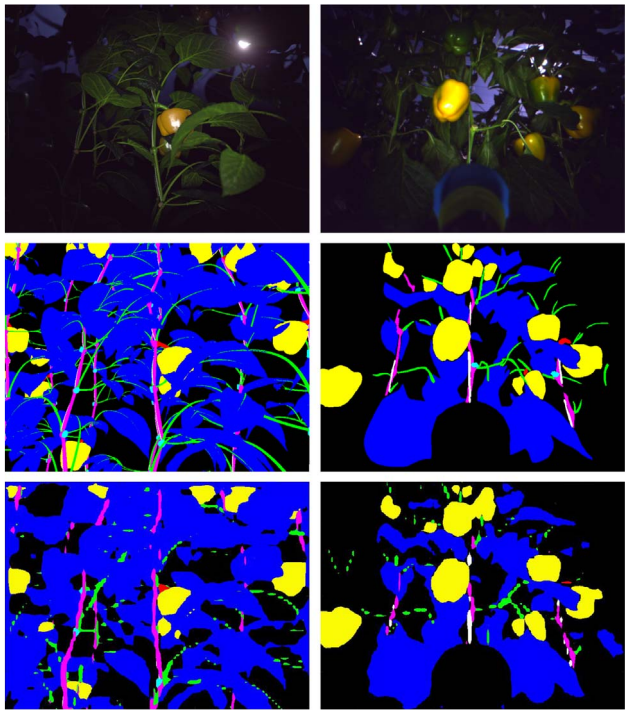

Fig. 7. Example segmentation results for synthetic test set from experiment II-A (left column) and empirical test set from experiment II-E (right column). Color images (top row), ground truth (middle row) and classification segmentation (bottom row) are shown. Class labels: ● background, ● leafs, ● peppers, ● peduncles, ● stems, ● shoots and leaf stems, ○ wires and ● cuts. (Forinterpretation of the references to colour, the reader is referred to the web version of this article.)

II-C When training a model using only a small empirical dataset, the performance of the most common classes approached the baseline performance of II-A, as can be seen in Fig. 8 and Appendix A. However, the model failed to discriminate the uncommon classes. It appeared the model only learned the most color discriminative classes.

II-D A model bootstrapped with a non-related dataset (PASCAL-VOC) that was fine-tuned with empirical data, resulted in increased performance on empirical data compared to the former experiments II-B and II-C where no fine-tuning was used. This implies fine-tuning on a bootstrapped network was beneficial. Any CNN network requires training time and a large dataset to converge to an effective feature distribution. Bootstrapping provides a stable starting point, from which the fine-tuning can quickly converge to a new optimum of the new dataset.

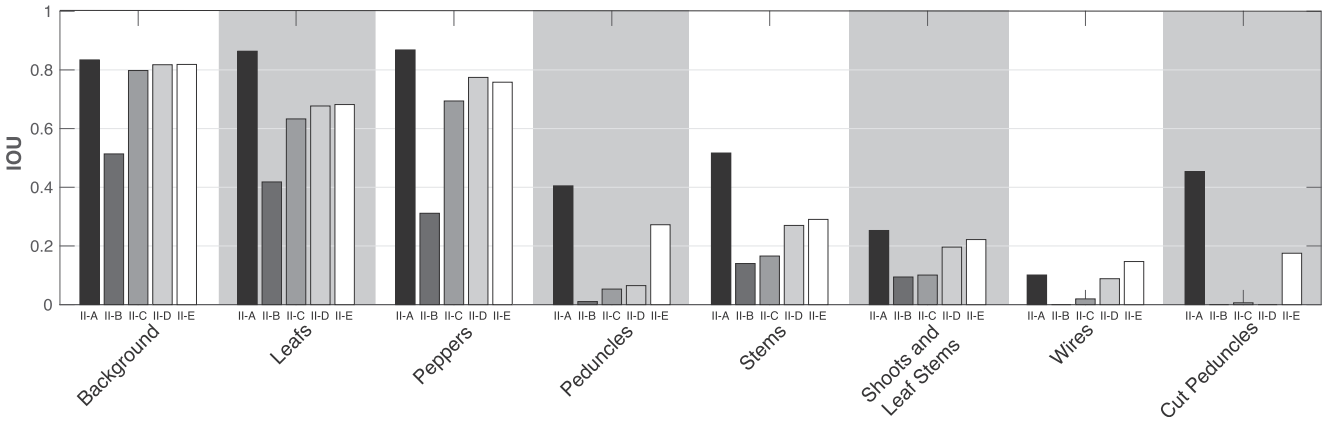

Fig. 8. Results of Experiment II. For each class, the mean test set IOU per class is displayed for each model II-A through II-E.

II-E The best IOU performance on empirical data and inclusion of all classes (see Appendix A) was achieved when bootstrapping on related synthetic data, confirming our hypothesis that a synthetic bootstrapping using synthetic data and fine-tuning with empirical data results in optimal learning.

When evaluating the results qualitatively in Fig. 7 we observe very high quality results in plant part recognition in both dataset types. Furthermore, although the IOU for some classes seems relatively low (IOU < 0.3) and segmentations were therefore not completely overlapping with the ground truth, we do observe good recall for each part nonetheless.

Note that for the empirical segmentation, parts in the image were detected that were not annotated manually in the ground truth due to dark regions, but were present in the image. Hence these were evaluated as false positives, although they would be true positives if human annotation was perfect. Hence this annotation bias resulted to a lower reported mean IOU.

For both segmentations it holds that elongated parts were not connected. This was likely due to the à *trous* algorithm that upscales a sparse low resolution input feature map (Chen et al., 2016a). Post processing with CRF might solve this issue, as investigated with Experiment VI in Section 9.

6. Experiment III

This experiment investigated the hypothesis of the positive correlation between segmentation performance and dataset size for our use-case, given a model with sufficient learnable parameters (Soekhoe et al., 2016; Zeiler and Fergus, 2014).

6.1. Methods

The DeepLab-V2 model was trained with logarithmically increasing synthetic dataset size, with image ranges III-A: $1\text{-}10^{1.0}$, III-B: $1\text{-}10^{1.5}$, III-C: $1\text{-}10^{2.0}$, III-D: $1\text{-}10^{2.5}$, III-E: $1\text{-}10^{3.0}$, III-F: $1\text{-}10^{3.5}$ and III-G $1\text{-}10^{3.942}$ (≈ 8750).

6.2. Results

In Fig. 9 the results for Experiment III are shown. For the underlying IOU distribution per class and over time, refer to Appendix A.

6.3. Discussion

Observing Fig. 9, the segmentation performance increased with dataset size though seemed to settle around 3,500 training images. As can be seen in Appendix A, the performance increase was mainly due to the rising correct classification of uncommon classes.

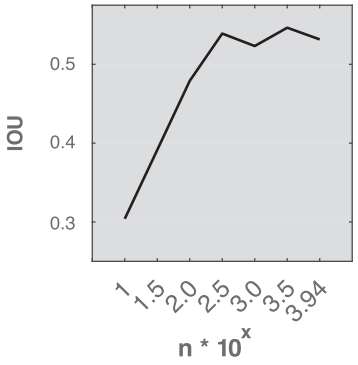

Fig. 9. Results of Experiment III: for an increasing synthetic dataset training size *n*, the mean IOU over the test set images and over all classes is displayed.

These results confirms our hypothesis that dataset size positively correlates with performance, in line with previous research of others (Zeiler and Fergus, 2014; Soekhoe et al., 2016). Additionally, it provided us the upper bound requirement for synthetic dataset size, as more synthetic training images do not further increase performance. However, the cause of the performance stabilisation might be twofold. First, our synthetic plant models may only capture a realistic but limited variation within the implemented plant parameter bounds. Hence, although each plant and scene was randomised (Barth et al., 2017), there should be an upper limit where generating new images merely adds redundant information. Therefore, if the synthetic models would improve on the variance similarity with the empirical situation, i.e. incorporating the wider range of plant parameters to increase diversity in the model, the performance might keep increasing further with dataset size.

Second, it might also be the case that the learning ability of the CNN model was saturated, meaning all weights could already be exploited and therefore the network was not able to absorb more information. A possible solution would be to raise the network's complexity, e.g. by increasing the number of feature maps and layers, whilst looking out for the overfitting pitfall using proper regularisation (Goodfellow et al., 2016).

7. Experiment IV

Until now the experiments investigated (i) which model would likely to be most suitable for our domain, (ii) whether it was possible to empirically fine-tune a synthetically bootstrapped model and (iii) how much synthetic data was required to bootstrap a model. As posed in Section 1.2, our main hypothesis is that only a small manually annotated dataset would be required for (ii). Experiment IV aims to further

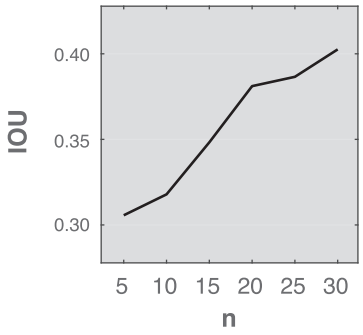

Fig. 10. Results of Experiment IV: for an increasing empirical fine-tuning dataset size *n*, the mean IOU over the test set images and over all classes is displayed.

specify the dataset size requirements as evidence for this hypothesis by evaluating how much empirical images are required to fine-tune a synthetically bootstrapped model.

7.1. Methods

We fine-tuned the DeepLab-V2 model that was synthetically bootstrapped with images 1–8750 with an increasing empirical dataset ranges of: IV-A: 1–5, IV-B: 1–10, IV-C: 1–15, IV-D: 1–20, IV-E: 1–25 and IV-F: 1–30. The performance was compared quantitatively by evaluating the mean IOU over the test set images and over all classes for each experiment.

7.2. Results

Results of Experiment IV is shown in Fig. 10.

7.3. Discussion

In Fig. 10 we observed that already with 5 empirical fine-tune images a reasonable performance can be achieved when a model is bootstrapped synthetically with related images (IOU = 0.306). This provides insight into the lower bound of manual annotated data required for fine-tuning, although this remains highly dependent on the IOU needed of a specific task. Results confirm our hypothesis only a small annotated dataset is sufficient for successful learning and meets our requirement of up to 30 annotated images.

From the figures in Appendix A we derived that fine-tuning settles in a minimum rapidly and furthermore overfitting was likely to occur when the training was not stopped prematurely.

Again the hypothesis was confirmed that an increase of dataset size improves performance. In our case a relative increase of 32% was achieved using 30 images as opposed to 5. Additionally, performance increase did not yet seem to settle, indicating that the small dataset did not yet cover all empirical variance and more empirical data is likely to further increase performance.

8. Experiment V

During the experiments we noted that common classes were better classified than uncommon classes.

A possible explanation could be in the nature of CNN weight learning. The weights of a convolutional neural network classifier move over the error landscape on the direction of the average gradient of the mini-batch. In the DeepLab architectures, a single training example consists of a cropped image of 300 × 300 pixels, assuming a batch size of 1. However, this image does not count as a single training example; instead the error of each pixel is used to determine the gradient. Given that the weights update in the average direction of the error, the gradient might therefore be biased towards common classes.

A possible solution is to normalise the per class error during the loss function computation. However, initial experiments where the loss was normalised by the number of each label present (as opposed to summing the loss) did not yield a significant difference in performance for any class.

We hypothesised further that the performance of individual part classes could be boosted by an aggregation of dedicated binary CNN models, one for each class. By applying this learning strategy, the error landscape is assumed to be simplified and therefore easier to learn with a bias to a single class.

8.1. Methods

Experiment V trained binary DeepLab-V2 models per plant part. To compare performance with the non-binary V2 model, the binary segmentations were aggregated by overlaying the output in descending order of IOU performance. The mean IOU over the test set images and over all classes of the aggregation was then compared to that of Experiment III-G, because those results were obtained using DeepLab-V2 without binary training.

8.2. Results

In Fig. 11 results are shown for the regular and binary models. In Appendix A the per class IOU performance can be observed.

8.3. Discussion

In Fig. 11 the results show that performance decreased relatively with 17% when training binary, as opposed to our hypothesis that this would improve results. Specifically the Cuts class underperformed significantly, as can be seen in Appendix A.

Although this experiment used a different training scheme, the underlying availability bias of each class remained equal. To cope with this difference, we suggest to balance the training data by cropping or masking the input proportionally to the class distributions presented in (Barth et al., 2017).

As each binary classifier was initialised with equal color normalisation parameters, based on the average color distribution over all classes, the data for each binary classifier was not zero-centered and normalised. However, attempts to normalise the data accordingly did not yield significant results.

9. Experiment VI

DeepLab models that were previously enhanced by adding fully

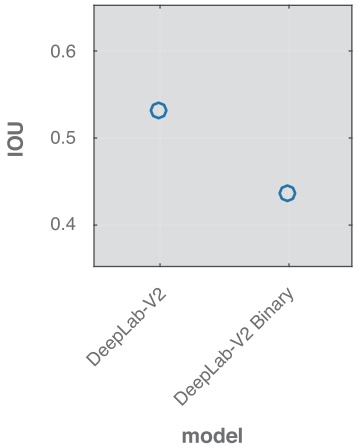

Fig. 11. Results of Experiment V. The mean IOU over the test set images and over all classes plotted for the DeepLab-V2 and aggregated binary DeepLab-V2 models.

conditional random fields (Krähenbühl and Koltun, 2011), showed improved segmentation performance for the PASCAL-VOC dataset (Chen et al., 2015, 2016a). In the previous Experiments I-V without CRF, the final classification for each pixel was determined by taking the maximum from all softmax class prediction layers. However, this approach disregards local and global label agreement, as similar labels tend to be clustered together and some labels co-occur more frequently than others. Applying a CRF can introduce local and global label agreement, usually resulting in a refinement of segmentation accuracy around the label edges.

9.1. Methods

Experiment VI optimised a CRF for the output of a synthetically trained model (1–8750) that was applied on synthetic data (8751–8800) (VI-A) and that was fine-tuned (1–30) and applied to empirical images (41–50) (VI-B). The optimisation of the CRF comprised of a selection of hyper-parameters, as described in (Krähenbühl and Koltun, 2011). To obtain the CRF parameters, we performed a coarse to fine grid search over a subset of possible parameters with values of [0 10 20 40] on the validation sets, similar to (Chen et al., 2015). The following values provided maximum IOU on the validation set of the synthetic data: $w_1 = 0.4$, $w_2 = 1.6$, $\sigma_\alpha = 0.1$, $\sigma_\beta = 0.04$ and $\sigma_\gamma = 9$ and on the empirical data: $w_1 = 0.2$, $w_2 = 0.5$, $\sigma_\alpha = 0.15$, $\sigma_\beta = 0.04$ and $\sigma_\gamma = 5$.

The performance was compared quantitatively by evaluating the mean IOU over the test set images and over all classes before and after applying the CRF. Furthermore, the final segmentations were assessed qualitatively to evaluate how the CRF influenced the segmentation.

9.2. Results

Post-processing the class probability maps output (see Fig. 5) of the DeepLab-V2 model using CRF resulted in an average IOU increase of 1.51% on the synthetic set but yielded marginal performance increase of 0.01% on the empirical set. To provide insight in qualitative improvement on the synthetic set, Fig. 12 displays an exemplary segmentation result with and without CRF as compared to the ground truth.

9.3. Discussion

Post-processing with CRF improved segmentation performance on the synthetic dataset both qualitatively as quantitatively with a relative +1.51%. The improvements were comparable to previously obtained results in other datasets (+3%) (Chen et al., 2015). Qualitatively we observed in Fig. 12 that disconnected circular class regions were connected and smoothed with sharp edges.

Unfortunately these results were not duplicated for the empirical dataset. Hence our hypothesis that the addition of CRFs improves local class segmentation boundaries only partially holds. Extended parameter search for the CRF did not provide better results. A possible explanation could be the small empirical test set size (10) as compared to the synthetic test set data size (50).

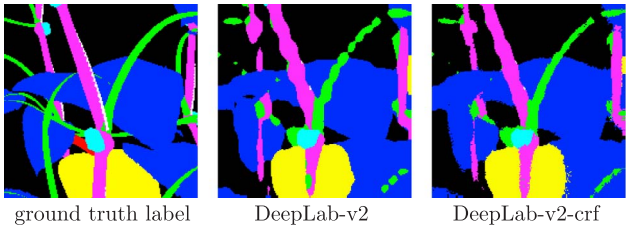

Fig. 12. Qualitative result of Experiment VI on a cropped image. The segmentation of the DeepLab-V2 model (middle) as compared to the ground truth (left) and with CRF post-processing (right). (For interpretation of the references to colour, the reader is referred to the web version of this article.)

10. Experiment VII

To test the DeepLab-V2 bootstrapped and fine-tuned model on generalisation power and robustness, we applied it to different but related datasets. Results provide insight in how applicable a model will be to new conditions. In turn this gives an estimate of the required additional annotation efforts when image acquisition conditions or scenes change.

10.1. Methods

In Experiment VII-A we deployed the model to datasets with equal empirical conditions but with different acquisition distances of 30, 20 and 10 cm from the crop. In Experiment VII-B we tested on a previous sweet-pepper image database obtained with different acquisition hardware. This dataset differed in artificial illumination, camera exposure settings, color calibration and crop season.

Due to the absence of a ground truth for these datasets, performance was compared qualitatively by evaluating the final segmentations. Specifically, we looked at the recognition of all classes and qualitatively at true and false positives.

10.2. Results

Example segmentation results are shown in Fig. 13.

10.3. Discussion

Evaluating the exemplary results in Fig. 13, we observed that the generalisation capability of an empirically fine-tuned CNN to other

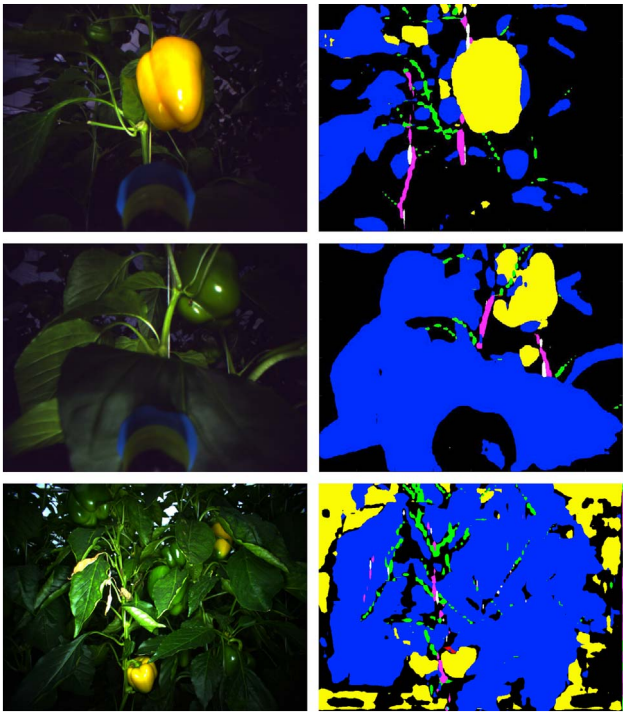

Fig. 13. Example segmentation results (right column) of the synthetically trained, empirically fine-tuned DeepLab-V2 model to images (left column) taken from 15 cm (row 1), 10 cm (row 2) and using different illumination hardware, exposure settings, color calibration and season (row 3). Class labels: ● background, ● leafs, ● peppers, ● peduncles, ● stems, ● shoots and leaf stems, ○ wires and ● cuts. (For interpretation of the references to colour, the reader is referred to the web version of this article.)

datasets was quite successful.

Images with similar hardware on closer distances were segmented similarly as the training set distance, although the number of false positive segmentations seemed to increase. Furthermore, not all classes seemed to be recognised (e.g. cuts).

Images from the different hardware, illumination conditions and color calibration and season, were still segmented fairly well, though mostly in the centered region of the image. Around the edges the probability for false positives seemed to increase, most probably due to relatively darker edges as compared with the empirical dataset on which it was trained.

The results suggest that empirical dataset fine-tune images do not necessarily have to be equal to the target image situation during deployment, as certain generalisation of the model can be expected. However, some performance degradation can be seen. Hence small additional manual annotation efforts are likely to be required. We partially confirmed our hypothesis that empirically fine-tuned CNNs can be applied robustly to related datasets.

11. General discussion

We compared our work to previously reported True Positive (TP) rates for plant part image segmentation in sweet-pepper (Bac et al., 2013) that used classification and regression trees (CART) on multi-spectral data. Performance differs as shown in Table 1. However, the previously reported measure in itself did not take into account other measures such as False Positives (FP). Furthermore, because TP rates can be maximised at the expense of increased FP rates, the overall performances of the methods were not directly comparable. However, as opposed to their conclusion (Bac et al., 2013), we hypothesise our segmentation results would be usable as reliable input for an obstacle map, because we obtained an empirical stem part IOU≈0.5. Furthermore, we observed qualitatively a low amount of false positive detections that might obfuscate a obstacle map.

In previous research by others, pursuing a comparable goal of segmenting plant species using synthetic image data (Cicco et al., 2016), similar results were obtained. Whilst training only on synthetic data and testing on empirical data, one of the models IOU performance for leaf class was 60.2, whereas when fine-tuned with empirical data the performance increased with 23% to 74.1. In our case, using the same training and testing methodology, the leaf class performance increased with 75% from an IOU of 0.4 to 0.7 (see Appendix A). However, we must note that the results were not directly comparable due to the difference in number of classes in each approach and the differing amount of empirical training data (30 vs 900). The authors do conclude similarly that a synthetic dataset can improve segmentation performance over the use of solely empirical images.

Although we aimed for a comprehensive set of experiments to obtain observations for our hypothesis and search for dataset requirements, we understand that our exploration was not exhaustive. However, we think our results show a clear direction of how important factors such as synthetic bootstrapping, architecture and dataset size and type influence part segmentation performance. Results showed state-of-the-art performance, given a minimal amount of manually annotated empirical fine-tuning data.

The results of Experiment III raised an important question we could not capture in our experiments. It remains unclear to what extent the complexity of the plant model and synthetic images influences the performance of bootstrapping. To answer this, the dataset complexity should be varied. However, this resided outside the scope of this paper and we suggest this as future research.

Related to this direction, would be informative to investigate the use of generative adversarial networks (GAN) to improve synthetic images towards the feature distribution of the empirical data (Shrivastava et al., 2016; Goodfellow et al., 2014).

Currently the DeepLab models do not differentiate between instances of parts. Future research on instance aware segmentation (Dai et al., 2016) could further improve the usability of the segmentations by discriminating between individual parts, for example by the MASK R-CNN architecture (He et al., 2017).

The difference between IOU performance of VOC and related dataset bootstrapping (II-D and II-E), was shown to be 15% relative and 0.05% absolute. It might be argued that related synthetic bootstrapping might not be worthwhile over bootstrapping with unrelated commonly available datasets. However, when we evaluated Fig. 8 and Appendix A, it shows that unrelated bootstrapping fails to recognise uncommon classes such as peduncles and cut peduncles. Our approach with related synthetic bootstrapping shows best performance and did recognise all classes.

Although there is a tradeoff measured in time investment between creating a synthetic dataset (Barth et al., 2017) and manually annotating additional empirical data, our results show that when empirical dataset size is a constraint and uncommon parts are required to be recalled, synthetic bootstrapping can provide a solution. Moreover, when a synthetic model is created formerly, it can be quickly used to generate synthetic data under a broad set of new application conditions whilst minimising the manual annotation requirement.

Table 1
True Positive rates of CART on part detection in hyperspectral sweet-pepper images and our CNN model.

	Leafs	Peppers	Peduncles	Stems
CART (Bac et al., 2013)	73.6	54.5	49.5	40.0
DeepLab-V2	78.5	34.5	78.6	21.6

12. Conclusion

In this paper we showed a methodology to reduce the current bottleneck of the reliance on manually annotated images that state-of-the-art machine learning requires. We provided evidence for our hypothesis that only a small manually annotated empirical fine-tuning dataset is still needed for optimal and successful empirical learning after bootstrapping a convolutional neural network (CNN) with related synthetic data. Our results show only 30 empirical training images were sufficient to obtain a mean IOU performance over all classes of 0.40. Furthermore, our method approached the synthetic baseline performance with a mean IOU over all classes of 0.53. Regarding our requirements, our method was (i) unique in ensuring the recognition of all classes, (ii) was optimal compared to other learning strategies and (iii) was evaluated qualitatively successful and had desired quantitative results for tasks such as part detection.

Experiments confirm our hypothesis that performance is positively correlated with dataset size both for the synthetic and empirical datasets, although there was an upper limit of synthetic data where performance stabilises. We suggested further research to further improve performance by increasing model complexity or synthetic data variance.

Of the VGG-16 model architectures that were investigated, the addition of à *trous* spatial pyramid pooling proved to be most effective. The post-processing by conditional random fields yielded a small performance boost in the synthetically trained networks, though failed to improve in the same amount on the empirical data. Training binary classifiers to improve uncommon class performance did not yield improved results. The generalisation capability to images under different conditions was demonstrated as feasible, though not equal in performance as when fine-tuned.

Our work provides the field of computer vision in agriculture a pioneering methodology for state-of-the-art segmentation performance, whilst simultaneously reducing the reliance on labour intensive manual annotations. Results are a key part in the next leap of robotization in

agriculture to keep up with the increasing demand of productivity and quality whilst decreasing the pressure on resources required.

Acknowledgements

The authors would like to thank prof. Dr. R. Howe and Dr. D. Perrin for their input of this research and making computing resources available. This research was partially funded by the European Commission in the Horizon2020 Programme (SWEEPER GA No. 644313).

Appendix A

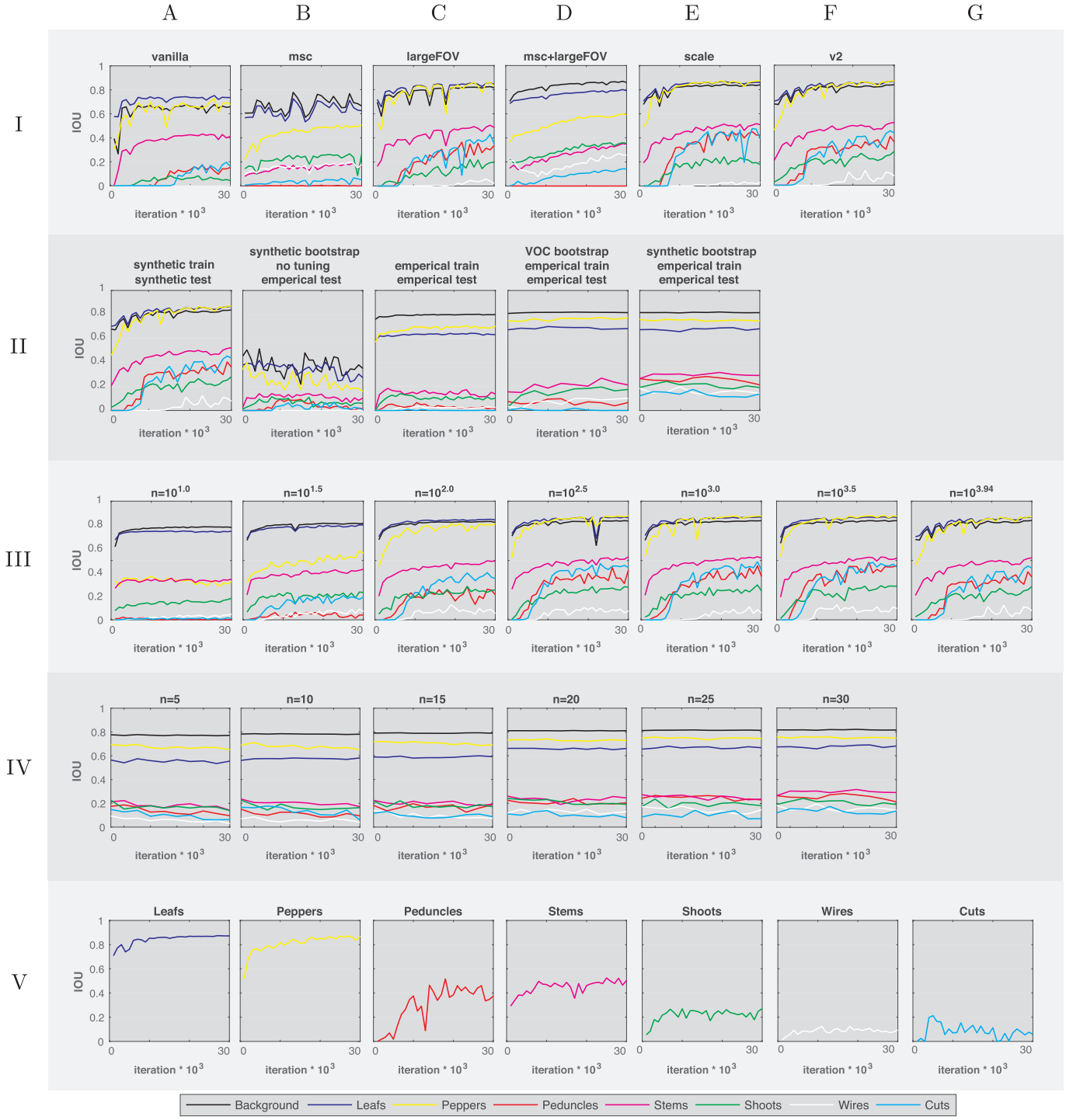

Fig. A.14. Detailed overview of IOU results of Experiments I through VII. For each model within those experiments A through F, the mean test set IOU, split by class, over the number of training iterations is displayed. Hence, figures show performance per class over time and the spread between all classes. (For interpretation of the references to colour, the reader is referred to the web version of this article.)

Appendix B. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.compag.2017.11.040.

References

Ahlin, K., Joffe, B., Hu, A.P., McMurray, G., Sadegh, N., 2016. Autonomous leaf picking using deep learning and visual-servoing. IFAC-PapersOnLine 49, 177–183. http://dx.doi.org/10.1016/j.ifacol.2016.10.033. 5th IFAC Conference on Sensing, Control and Automation Technologies for Agriculture AGRICONTROL 2016. http://www.sciencedirect.com/science/article/pii/S2405896316315968.

Araus, J.L., Cairns, J.E., 2014. Field high-throughput phenotyping: the new crop breeding frontier. Trends Plant Sci. 19, 52–61. http://dx.doi.org/10.1016/j.tplants.2013.09.008. http://www.sciencedirect.com/science/article/pii/S1360138513001994.

Bac, C., Hemming, J., van Henten, E., 2013. Robust pixel-based classification of obstacles for robotic harvesting of sweet-pepper. Comput. Electron. Agric. 96, 148–162. http://dx.doi.org/10.1016/j.compag.2013.05.004. http://www.sciencedirect.com/science/article/pii/S0168169913001099.

Bac, C., Hemming, J., van Henten, E., 2014a. Stem localization of sweet-pepper plants using the support wire as a visual cue. Comput. Electron. Agric. 105, 111–120. http://dx.doi.org/10.1016/j.compag.2014.04.011. http://www.sciencedirect.com/science/article/pii/S0168169914000933.

Bac, C.W., van Henten, E.J., Hemming, J., Edan, Y., 2014b. Harvesting robots for high-value crops: state-of-the-art review and challenges ahead. J. Field Robot. 31, 888–911. http://dx.doi.org/10.1002/rob.21525.

Bac, C.W., Roorda, T., Reshef, R., Berman, S., Hemming, J., van Henten, E.J., 2016. Analysis of a motion planning problem for sweet-pepper harvesting in a dense obstacle environment. Biosyst. Eng. 146, 85–97. http://dx.doi.org/10.1016/j.biosystemseng.2015.07.004. special Issue: Advances in Robotic Agriculture for Crops. http://www.sciencedirect.com/science/article/pii/S1537511015001191.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: a deep convolutional encoder-decoder architecture for scene segmentation. IEEE Trans. Pattern Anal. Mach. Intell. PP 1. http://dx.doi.org/10.1109/TPAMI.2016.2644615.

Banko, M., Brill, E., 2001. Scaling to very very large corpora for natural language disambiguation. In: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 26–33. http://dx.doi.org/10.3115/1073012.1073017.

Barth, R., IJsselmuiden, J., Hemming, J., van Henten, E., 2017. Data synthesis methods for semantic segmentation in agriculture: a capsicum annum dataset. J. Comput. Electron. Agric. http://dx.doi.org/10.1016/j.compag.2017.12.001. (Submitted for publication).

Bengio, Y., 2009. Learning deep architectures for ai. Found. Trends Mach. Learn. 2, 1–127. http://dx.doi.org/10.1561/2200000006.

Bengio, Y., 2011. Deep learning of representations for unsupervised and transfer learning. In: Proceedings of the 2011 International Conference on Unsupervised and Transfer Learning Workshop – Volume 27, JMLR.org., pp. 17–37. <http://dl.acm.org/citation.cfm?id=3045796.3045800>.

Bengio, Y., 2012. Practical recommendations for gradient-based training of deep architectures. CoRR abs/1206.5533. http://arxiv.org/abs/1206.5533.

Bengio, Y., Bastien, F., Bergeron, A., Boulanger-lewandowski, N., Breuel, T.M., Chherawala, Y., Cisse, M., Ct, M., Erhan, D., Eustache, J., Glorot, X., Muller, X., Lebeuf, S.P., Pascanu, R., Rifai, S., Savard, F., Sicard, G., 2011. Deep learners benefit more from out-of-distribution examples. In: Gordon, G.J., Dunson, D.B. (Eds.), Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS-11), Journal of Machine Learning Research - Workshop and Conference Proceedings, pp. 164–172.

Brants, T., Popat, A.C., Xu, P., Och, F.J., Dean, J., 2007. Large language models in machine translation. In: Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 858–867.

Caruana, R., 1995. Learning many related tasks at the same time with backpropagation. In: Advances in Neural Information Processing Systems 7, Morgan Kaufmann, pp. 657–664.

Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2015. Semantic image segmentation with deep convolutional nets and fully connected crfs. In: ICLR.

Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2016a. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. Also avaialble at: arXiv:1606.00915.

Chen, L.C., Yang, Y., Wang, J., Xu, W., Yuille, A.L., 2016b. Attention to scale: scale-aware semantic image segmentation. In: CVPR.

Chen, X., Mottaghi, R., Liu, X., Fidler, S., Urtasun, R., Yuille, A., 2014. Detect what you can: Detecting and representing objects using holistic models and body parts. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Cicco, M.D., Potena, C., Grisetti, G., Pretto, A., 2016. Automatic model based dataset generation for fast and accurate crop and weeds detection. CoRR abs/1612.03019. http://arxiv.org/abs/1612.03019.

Cordier, N., Delingette, H., Le, M., Ayache, N., 2016. Extended modality propagation: image synthesis of pathological cases. IEEE Trans. Med. Imag. PP 1. http://dx.doi.org/10.1109/TMI.2016.2589760.

Dai, J., He, K., Sun, J., 2016. Instance-aware semantic segmentation via multi-task network cascades. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Dittrich, F., Woern, H., Sharma, V., Yayilgan, S., 2014. Pixelwise object class segmentation based on synthetic data using an optimized training strategy. In: 2014 First International Conference on Networks Soft Computing (ICNSC), pp. 388–394. http://dx.doi.org/10.1109/CNSC.2014.6906671.

Erhan, D., Bengio, Y., Courville, A., Manzagol, P.A., Vincent, P., Bengio, S., 2010. Why does unsupervised pre-training help deep learning? J. Mach. Learn. Res. 11, 625–660. http://dl.acm.org/citation.cfm?id=1756006.1756025.

Everingham, M., Eslami, S.M., Gool, L., Williams, C.K., Winn, J., Zisserman, A., 2015. The pascal visual object classes challenge: a retrospective. Int. J. Comput. Vision 111, 98–136. http://dx.doi.org/10.1007/s11263-014-0733-5.

Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A., 2010a. The pascal visual object classes (voc) challenge. Int. J. Comput. Vis., vol. 88, pp. 303–338.

Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A., 2010b. The pascal visual object classes (voc) challenge. Int. J. Comput. Vis., vol. 88, pp. 303–338.

Farabet, C., Couprie, C., Najman, L., LeCun, Y., 2013. Learning hierarchical features for scene labeling. IEEE Trans. Pattern Anal. Mach. Intell. 35, 1915–1929. http://dx.doi.org/10.1109/TPAMI.2012.231.

Felzenszwalb, P.F., Huttenlocher, D.P., 2000. Efficient matching of pictorial structures. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No.PR00662), pp. 66–73, vol. 2. http://dx.doi.org/10.1109/CVPR.2000.854739.

Felzenszwalb, P.F., Huttenlocher, D.P., 2005. Pictorial structures for object recognition. Int. J. Comput. Vision 61, 55–79. http://dx.doi.org/10.1023/B:VISI.0000042934.15159.49.

Fulkerson, B., Vedaldi, A., Soatto, S., 2009. Class segmentation and object localization with superpixel neighborhoods. In: Proceedings of the International Conference on Computer Vision (ICCV).

Gongal, A., Amatya, S., Karkee, M., Zhang, Q., Lewis, K., 2015. Sensors and systems for fruit detection and localization: a review. Comput. Electron. Agric. 116, 8–19.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press. <http://www.deeplearningbook.org>.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 27. Curran Associates, Inc., pp. 2672–2680. <http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf>.

Hattori, H., Boddeti, V.N., Kitani, K., Kanade, T., 2015. Learning scene-specific pedestrian detectors without real data. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3819–3827. http://dx.doi.org/10.1109/CVPR.2015.7299006.

He, H., Garcia, E.A., 2009. Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. 21, 1263–1284. http://dx.doi.org/10.1109/TKDE.2008.239.

He, K., Gkioxari, G., Dollár, P., Girshick, R.B., 2017. Mask R-CNN. CoRR abs/1703.06870. http://arxiv.org/abs/1703.06870.

He, K., Zhang, X., Ren, S., Sun, J., 2014. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. Springer International Publishing, Cham. http://dx.doi.org/10.1007/978-3-319-10578-9_23.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: surpassing human-level performance on imagenet classification. CoRR abs/1502.01852. http://arxiv.org/abs/1502.01852.

van der Heijden, G., Song, Y., Horgan, G., Polder, G., Dieleman, A., Bink, M., Palloix, A., van Eeuwijk, F., Glasbey, C., 2012. Spicy: towards automated phenotyping of large pepper plants in the greenhouse. Funct. Plant Biol. 39, 870–877. http://dx.doi.org/10.1071/FP12019.

Hemming, J., Rath, T., 2001. Paprecision agriculture: computer-vision-based weed identification under field conditions using controlled lighting. J. Agric. Eng. Res. 78, 233–243. http://dx.doi.org/10.1006/jaer.2000.0639. http://www.sciencedirect.com/science/article/pii/S0021863400906395.

Henten, E.V., Slot, D.V., Hol, C., Willigenburg, L.V., 2009. Optimal manipulator design for a cucumber harvesting robot. Comput. Electron. Agric. 65, 247–257. http://dx.doi.org/10.1016/j.compag.2008.11.004. http://www.sciencedirect.com/science/article/pii/S0168169908002238.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T., 2014. Caffe: Convolutional architecture for fast feature embedding. Also Available At: arXiv:1408.5093.

Johnson, M., Shotton, J., Cipolla, R., 2013. Semantic Texton Forests for Image Categorization and Segmentation. Springer London, London. http://dx.doi.org/10.1007/978-1-4471-4929-3_15. pp. 211–227.

Kavasidis, I., Palazzo, S., Salvo, R.D., Giordano, D., Spampinato, C., 2014. An innovative web-based collaborative platform for video annotation. Multimedia Tools Appl. 70, 413–432. http://dx.doi.org/10.1007/s11042-013-1419-7.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. CoRR abs/1412.6980. http://arxiv.org/abs/1412.6980.

Kondaveeti, H.K., 2016. Synthetic isar images of aircrafts. http://dx.doi.org/10.5281/zenodo.48002.

Krähenbühl, P., Koltun, V., 2011. Efficient inference in fully connected crfs with gaussian edge potentials. In: NIPS.

Learned-Miller, E., Huang, G.B., RoyChowdhury, A., Li, H., Hua, G., 2016. Labeled Faces in the Wild: A Survey. Springer International Publishing, Cham. http://dx.doi.org/10.

1007/978-3-319-25958-1_8. pp. 189–248.
Li, Y., Cao, Z., Lu, H., Xiao, Y., Zhu, Y., Cremers, A.B., 2016. In-field cotton detection via region-based semantic image segmentation. Comput. Electron. Agric. 127, 475–486. http://dx.doi.org/10.1016/j.compag.2016.07.006. http://www.sciencedirect.com/science/article/pii/S016816991630480X.
Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
Lu, J., Hu, J., Zhao, G., Mei, F., Zhang, C., 2017. An in-field automatic wheat disease diagnosis system. Comput. Electron. Agric. 142, 369–379. http://dx.doi.org/10.1016/j.compag.2017.09.012. http://www.sciencedirect.com/science/article/pii/S0168169917305999.
Mallat, S., 1999. A Wavelet Tour of Signal Processing. second ed. Academic Press, San Diego.
Milioto, A., Lottes, P., Stachniss, C., 2017a. Real-time blob-wise sugar beets vs weeds classification for monitoring fields using convolutional neural networks. ISPRS Annals Photogramm., Rem. Sens. Spatial Inform. Sci., IV-2/W3, pp. 41–48. <https://www.isprs-ann-photogramm-remote-sens-spatial-inf-sci.net/IV-2-W3/41/2017/>. http://dx.doi.org/10.5194/isprs-annals-IV-2-W3-41-2017.
Milioto, A., Lottes, P., Stachniss, C., 2017b. Real-time semantic segmentation of crop and weed for precision agriculture robots leveraging background knowledge in cnns.
Mishkin, D., Matas, J., 2015. All you need is a good init. CoRR abs/1511.06422. http://arxiv.org/abs/1511.06422.
Mostajabi, M., Yadollahpour, P., Shakhnarovich, G., 2014. Feedforward semantic segmentation with zoom-out features. CoRR abs/1412.0774. http://arxiv.org/abs/1412.0774.
Najafabadi, M.M., Villanustre, F., Khoshgoftaar, T.M., Seliya, N., Wald, R., Muharemagic, E., 2015. Deep learning applications and challenges in big data analytics. J. Big Data 2, 1. http://dx.doi.org/10.1186/s40537-014-0007-7.
Oberti, R., Marchi, M., Tirelli, P., Calcante, A., Iriti, M., Hoevar, M., Baur, J., Pfaff, J., Ulbrich, H., 2013. Selective spraying of grapevines diseases by a modular agricultural robot 44, 149–153.
Papandreou, G., Chen, L.C., Murphy, K., Yuille, A.L., 2015. Weakly- and semi-supervised learning of a dcnn for semantic image segmentation. In: ICCV.
Phadikar, S., Sil, J., 2008. Rice disease identification using pattern recognition techniques. In: 2008 11th International Conference on Computer and Information Technology, pp. 420–423. http://dx.doi.org/10.1109/ICCITECHN.2008.4803079.
Polder, G., van der Heijden, G.W., van Doorn, J., Baltissen, T.A., 2014. Automatic detection of tulip breaking virus (tbv) intulip fields using machine vision. Biosyst. Eng. 117, 35–42. http://dx.doi.org/10.1016/j.biosystemseng.2013.05.010. Image Analysis in Agriculture. http://www.sciencedirect.com/science/article/pii/S1537511013000883.
Pont-Tuset, J., Arbeláez, P., Barron, J., Marques, F., Malik, J., 2015. Multiscale combinatorial grouping for image segmentation and object proposal generation. Also Available at: arXiv:1503.00848.
Pound, M.P., Atkinson, J.A., Wells, D.M., Pridmore, T.P., French, A.P., 2017. Deep learning for multi-task plant phenotyping. bioRxiv. https://www.biorxiv.org/content/early/2017/10/17/204552, http://dx.doi.org/10.1101/204552, Also Available at: arXiv:https://www.biorxiv.org/content/early/2017/10/17/204552.full.pdf.
Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A., 2016. The SYNTHIA Dataset: A large collection of synthetic images for semantic segmentation of urban scenes.
Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T., 2008. Labelme: a database and web-based tool for image annotation. Int. J. Comput. Vision 77, 157–173. http://dx.doi.org/10.1007/s11263-007-0090-8.
Sa, I., Lehnert, C., English, A., McCool, C., Dayoub, F., Upcroft, B., Perez, T., 2017. Peduncle detection of sweet pepper for autonomous crop harvesting combined color and 3-d information. IEEE Robot. Autom. Lett. 2, 765–772. http://dx.doi.org/10.1109/LRA.2017.2651952.
Shapiro, D., 2016. Accelerating the race to autonomous cars. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA. pp. 415–415. http://dx.doi.org/10.1145/2939672.2945360.
Shelhamer, E., Long, J., Darrell, T., 2016. Fully convolutional networks for semantic segmentation. CoRR abs/1605.06211.
Shotton, J., Girshick, R., Fitzgibbon, A., Sharp, T., Cook, M., Finocchio, M., Moore, R., Kohli, P., Criminisi, A., Kipman, A., Blake, A., 2013. Efficient Human Pose Estimation from Single Depth Images. Springer London, London. http://dx.doi.org/10.1007/978-1-4471-4929-3_13. pp. 175–192.
Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., Webb, R., 2016. Learning from simulated and unsupervised images through adversarial training. CoRR abs/1612.07828. http://arxiv.org/abs/1612.07828.
Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556.
Soekhoe, D., van der Putten, P., Plaat, A., 2016. On the Impact of Data Set Size in Transfer Learning Using Deep Neural Networks. Springer International Publishing, Cham. http://dx.doi.org/10.1007/978-3-319-46349-0_5.
de Soto, M.G., Emmi, L., Perez-Ruiz, M., Aguera, J., de Santos, P.G., 2016. Autonomous systems for precise spraying evaluation of a robotised patch sprayer. Biosyst. Eng. 146, 165–182. http://dx.doi.org/10.1016/j.biosystemseng.2015.12.018. special Issue: Advances in Robotic Agriculture for Crops. http://www.sciencedirect.com/science/article/pii/S153751101500197X.
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: A simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15, 1929–1958. http://jmlr.org/papers/v15/srivastava14a.html.
Tsogkas, S., Kokkinos, I., Papandreou, G., Vedaldi, A., 2015. Semantic part segmentation with deep learning. CoRR abs/1505.02438. http://arxiv.org/abs/1505.02438.
Uijlings, J.R.R., van de Sande, K.E.A., Gevers, T., Smeulders, A.W.M., 2013. Selective search for object recognition. Int. J. Comput. Vis. 104, 154–171. https://ivi.fnwi.uva.nl/isis/publications/2013/UijlingsIJCV2013.
Vijayarangan, S., Sodhi, P., Kini, P., Bourne, J., Du, S., Sun, H., Poczos, B., Apostolopoulos, D.D., Wettergreen, D., 2017. High-throughput robotic phenotyping of energy sorghum crops. In: Field and Service Robotics, Springer-Verlag.
Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S., 2011. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report.
Wang, J., Yuille, A.L., 2015. Semantic part segmentation using compositional model combining shape and appearance. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
Wang, L., Shi, J., Song, G., Shen, I.f., 2007. Object Detection Combining Recognition and Segmentation. Springer Berlin Heidelberg, Berlin, Heidelberg. http://dx.doi.org/10.1007/978-3-540-76386-4_17. pp. 189–199.
Wehrens, R., 2010. Self-Organising Maps for Image Segmentation. Springer Berlin Heidelberg, Berlin, Heidelberg. http://dx.doi.org/10.1007/978-3-642-01044-6_34. pp. 373–383.
Wu, Z., Shen, C., van den Hengel, A., 2016. Wider or deeper: Revisiting the resnet model for visual recognition. CoRR abs/1611.10080. http://arxiv.org/abs/1611.10080.
Zeiler, M.D., Fergus, R., 2014. Visualizing and Understanding Convolutional Networks. Springer International Publishing, Cham. http://dx.doi.org/10.1007/978-3-319-10590-1_53. pp. 818–833.
Zeng, A., Yu, K., Song, S., Suo, D. Jr., E.W., Rodriguez, A., Xiao, J., 2016. Multi-view self-supervised deep learning for 6d pose estimation in the amazon picking challenge. CoRR abs/1609.09475. http://arxiv.org/abs/1609.09475.
Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2016. Pyramid Scene Parsing Network. ArXiv e-prints arXiv:1612.01105.
Zhu, H., Meng, F., Cai, J., Lu, S., 2016. Beyond pixels: a comprehensive survey from bottom-up to semantic image segmentation and cosegmentation. J. Visual Commun. Image Represent. 34, 12–27. http://dx.doi.org/10.1016/j.jvcir.2015.10.012. http://www.sciencedirect.com/science/article/pii/S1047320315002035.