Computers and Electronics in Agriculture 173 (2020) 105378

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

# Optimising realism of synthetic images using cycle generative adversarial networks for improved part segmentation

R. Barth[a,*], 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, Droevendaalsesteeg 1, 6708 PB Wageningen, the Netherlands

ARTICLE INFO

*Keywords:*
Synthetic dataset
Semantic segmentation
3D modelling
Agriculture
Robotics

ABSTRACT

In this paper we report on improving part segmentation performance for robotic vision using convolutional neural networks by optimising the visual realism of synthetic agricultural images. In Part I, a cycle consistent generative adversarial network was applied to synthetic and empirical images with the objective to generate more realistic synthetic images by translating them to the empirical domain. We hypothesise that plant part image features (e.g. color, texture) become more similar to the empirical domain after translation of the synthetic images. Results confirm this with an improved mean color distribution correlation with the empirical data prior of 0.62 and post translation of 0.90. Furthermore, the mean image features of contrast, homogeneity, energy and entropy moved closer to the empirical mean, post translation. In Part II, 7 experiments were performed using convolutional neural networks with different combinations of synthetic, synthetic translated to empirical and empirical images. We hypothesise that the translated images can be used for (i) improved learning of empirical images, and (ii) that learning without any fine-tuning with empirical images is improved by bootstrapping with translated images over bootstrapping with synthetic images.

Results confirm our hypotheses in Part II. First a maximum intersection-over-union performance was achieved of 0.52 when bootstrapping with translated images and fine-tuning with empirical images; an 8% increase compared to only using synthetic images. Second, training without any empirical fine-tuning resulted in an average IOU of 0.31; a 55% performance increase over previous methods that only used synthetic images. The key contribution of this paper to robotic vision is to provide supporting evidence that domain adaptation can be successfully used to translate and improve synthetic data to the real empirical domain that results in improved segmentation learning whilst lowering the dependency on manually annotated data.

## 1. Introduction

A key success factor of robotics performance is a robust underlying perception methodology that can distinguish and localise object parts (Bac et al., 2013; Gongal et al., 2015; Bac et al., 2014). In order to train state-of-the-art machine learning methods that can achieve this feat, large annotated empirical image datasets remain required. Synthetic images can help bootstrapping such methods in order to reduce the required amount of annotated empirical data (Barth et al., 2017a). However, a gap in realism remains between the modelled synthetic images and the empirical ones, plausibly restraining synthetic bootstrapping performance.

The long term objective of our research is to improve plant part segmentation performance, applied in the field of agricultural robotics. Previous work performed synthetically bootstrapping deep convolutional neural networks (CNN) (Barth et al., 2017a). In this paper we report on optimising the realism of rendered synthetic images modelled from empirical photographical data (Barth et al., 2017b) that was used in our previous work. We first hypothesise that the dissimilarity between synthetic and empirical images can be qualitatively and quantitatively reduced using unpaired image-to-image translation by cycle-consistent adversarial networks (Cycle-GAN) (Zhu et al., 2017). Furthermore, we secondly hypothesise that the synthetic images translated to the empirical domain can be used for improved learning of empirical images, potentially further closing the performance gap that remained previously when bootstrapping only with synthetic data (Barth et al., 2017a). Additionally, our third hypothesis is that without any fine-tuning with empirical images, improved learning of empirical images can be achieved using only translated images as opposed to using only synthetic images.

* Corresponding author.
*E-mail addresses:* ruud.barth@gmail.com (R. Barth), jochen.hemming@wur.nl (J. Hemming), eldert.vanhenten@wur.nl (E.J. Van Henten).

https://doi.org/10.1016/j.compag.2020.105378
Received 17 October 2019; Received in revised form 13 March 2020; Accepted 18 March 2020

The key contribution of this paper to robotic vision is to provide supporting evidence that domain adaptation can be successfully used to translate and improve synthetic data to the real empirical domain that results in improved segmentation learning whilst lowering the dependency on manually annotated data.

The contributions of this work to the field of domain adaptation are twofold, both aimed to verify the translation performance. First, a method is provided to analyse class color distributions and images features such as contrast, homogeneity, energy and entropy. Second, an approach is given to exhaustively analyse the effect of different training paradigms and dataset compositions and sizes on classification performance. Together, these methods support confirming the relevancy and effectiveness of synthetic data combined with domain adaptation for learning.

### 1.1. Theoretical background

Convolutional neural networks recently have shown state-of-the-art performance on many image segmentation tasks (Chen et al., 2017; Long et al., 2015b; Chen et al., 2018). However, CNNs require large annotated datasets on a per-pixel level in order to successfully train the large number of free parameters of the deep network. Moreover, in agriculture the high amount of image variety due to a wide range of species, illumination conditions and morphological seasonal growth differences, leads to an increased annotated dataset size dependency. Satisfying this requirement can quickly become a bottleneck for learning.

One solution is to bootstrap CNNs with synthetic images including automatically computed ground truths (Dittrich et al., 2014; Ros et al., 2016). Consequently, the bootstrapped network can be fine-tuned with a small set of empirical images, which can result in increased performance over methods without synthetic bootstrapping (Barth et al., 2017a).

Previously we have shown methods to create such a synthetic dataset by realistically rendering 3D modelled plants (Barth et al., 2017b). Despite intensive manual optimisation for geometry, color and textures, we have shown that a discrepancy remains between the synthetic and empirical images. Although this dataset can be used for successful synthetic bootstrapping and improved empirical learning, there remained a difference between the achieved performance and the theoretical optimal performance (Barth et al., 2017a).

Recently, the advent of generative adversarial networks (GAN) introduced another method of image data generation (Goodfellow et al., 2014). In GANs two deep convolutional neural networks are trained simultaneously and adversarially: a generative model $G$ and a discriminative model $D$. The generative model's goal is to capture the feature distribution of a dataset by learning to generate images thereof from latent variables (e.g. random noise vectors). The discriminative model in turn evaluates to what extent the generated image is a true member of the dataset. In other words, model $G$ is optimised to trick model $D$ while model $D$ is optimising to not get fooled by model $G$. As both models can be implemented as CNNs, the error can be back-propagated to minimise the loss of both models simultaneously. The result after training is a model $G$ that can generate new random images highly similar to the learned dataset. This method is useful if one wants to generate more similar images from the same domain. Given that this does not provide a corresponding ground truth, this method was not pursued for this paper, although as we'll see later, it does provide a fundamental building block towards the method that was used.

In later approaches, GANs were conditioned with an additional input image from another domain (Isola et al., 2017), forming an image pair that had some relation with each other (e.g. a color image and its label or class mapping). The generator was tasked with image-to-image translation to create a coherent image (e.g. color) from a corresponding pair image (e.g. label map). The discriminator's goal is then to evaluate if input pairs are either real or generated. The loss can then be fed back to both the discriminator and generator to improve on their tasks. The result after training is a generator $G$ that can translate images from one domain X (e.g. color images) to images in another domain Y (e.g. label maps) or more formally notated as $G: X \rightarrow Y$. Given that this method does not provide additional novel training pairs, we also did not pursue this method for this paper, although again this methodology provides a useful building block for our work. A requirement for image-to-image translation using conditional GANs, is that images from both domains are geometrically paired. For our objective of translating images from the synthetic domain to the empirical domain, this requirement was not met because images from both domains did not geometrically correspond one-to-one.

A recent approach aimed to dissolve this paired geometry requirement by investigating unpaired image-to-image translation (Zhu et al., 2017). In cycle-consistent adversarial networks (Cycle-GAN), a mapping $G: X \rightarrow Y$ is learned whilst at the same time also the inverse mapping $F: Y \rightarrow X$ is learned. Both domains $X$ and $Y$ have corresponding discriminators $D_X$ and $D_Y$. Hence, $D_X$ ensures $G$ to translate $X$ similar to $Y$ whilst $D_Y$ tries to safeguard a preferably indistinguishable conversion of $Y$ to $X$.

However since the domains are unpaired, the translation at this point does not guarantee that an individual image $x \in X$ is mapped to a geometrically similar image in domain $Y$ (or vice versa $y \in Y$ to $X$). This is because there are boundless mappings from $x$ that result in the same target distribution of $Y$. Therefore the mapping needs to be constrained in a way such that the original geometry is maintained.

To achieve that, a cycle consistency loss was added to further regularise the learning (Zhu et al., 2017). Given a sample $x \in X$ and $y \in Y$, a loss was added to the optimisation such that $F(G(x)) \approx x$ and $G(F(y)) \approx y$. Hence, the learning was therefore constrained by the intuition that if an input image is translated from one domain to the other and then back again, an image should result similar to the original input. This similarity is captured by the cycle consistency loss, which forces the generators $G$ and $F$ to achieve unpaired geometrically consistent image-to-image translation from one domain to the other and vice versa.

In Fig. 1 a schematic is shown of this learning process. Note that this method was pursued for this paper, because it allows to create a large dataset of images in the empirical domain. Furthermore, the key utility lies in the image pair P, in which the ground truth class mapping from the synthetic images could also be used for the translated synthetic images to the empirical domain. Moreover, the method does not require any annotations of empirical images.

### 1.2. Related work

When it first became clear that the new generation of classifiers, e.g. convolutional neural networks, could cope and benefit from large datasets, the use of synthetic image generation was investigated (Shotton et al., 2011; Wood et al., 2016; Qiu and Yuille, 2016; Richter et al., 2016a,b). Although pre-training with images from the synthetic domain was beneficial, often a performance gap remained compared to training with equally large datasets of the real empirical domain. This gap was thought to be caused by a dataset bias (Quionero-Candela et al., 2019), notably in a difference in realism, that was still to large for the learning algorithms to bridge.

Approaches to reduce this gap, intersect with the emerged specialised field of domain adaptation (Saenko et al., 2010) as a part of the larger transfer learning problem (Csurka, 2017), where labeled data in related source domains are used to learn new or unlabelled data in a target domain. The differences between domains may include visual features such as lighting or color arrangements, but also camera properties and poses. Additionally, some classes may be absent in one set or have a different distribution in another.

In the beginning, domain adaptation aligned features from both domains using correlation distance (Sun and Saenko, 2016) or maximum mean discrepancy (Long et al., 2015a), but these methods were

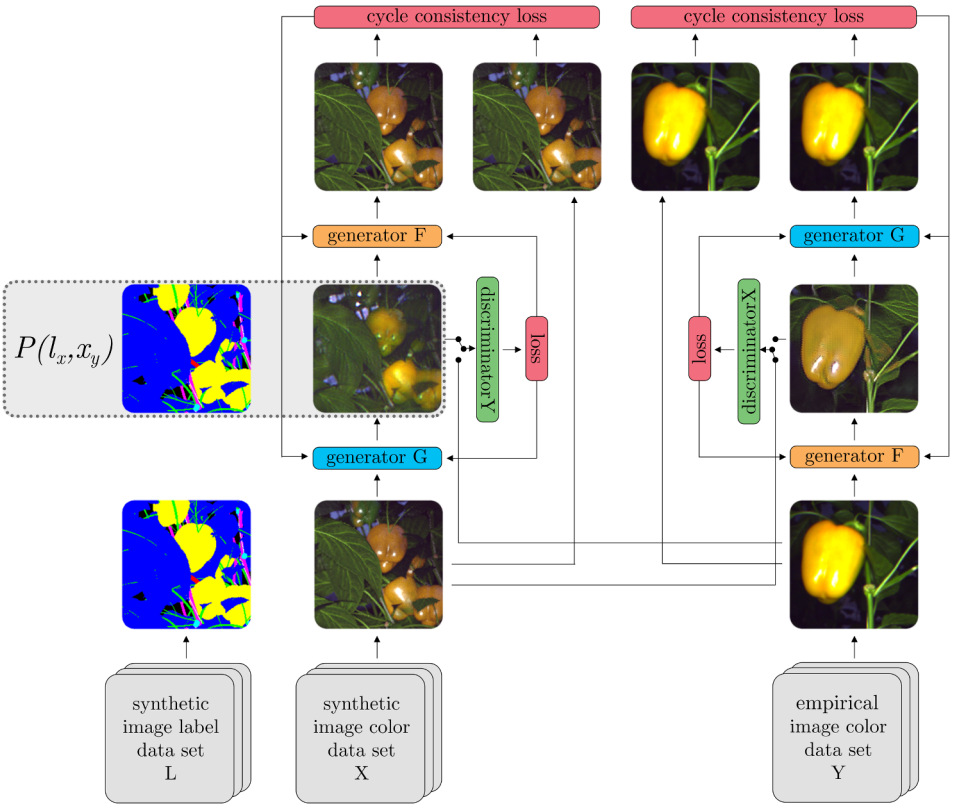


**Fig. 1.** Learning schematic of a cycle generative adversarial network. In each learning step, generator G receives an image from domain X and generator F receives an image from domain Y. Each generator is trained to transform the input image to the other domain. A discriminator Y and discriminator X for each corresponding domain is trained to distinguish between generated and original domain images. From those first set of generated images, the opposing generator then synthesizes the second set of images back to its original domain (which ideally should result in the original domain image). A cycle consistency loss is then calculated by comparing the second set of images with the initial input image. The loss of both discriminators and cycle consistency is fed back to both generators for learning. In this example, each generator learns to synthesise an image to the opposing domain, whilst remaining geometrically consistent. This example was pursued in this paper to obtain image pair P, consisting of the label $l_x$ that corresponds to $x_y$; the translated image from domain X to Y. Note that pair P is not input of Discriminator Y, only the color image from Generator G is.

limited by semantic consistency. Approaches to overcome this used generative adversarial methods and worked on the raw pixel space (Shrivastava et al., 2017). For example by using GANs with synthetic images as inputs instead of random vectors while preserving the annotation information, achieving state-of-the-art results for pose or gaze classification (Shrivastava et al., 2017). Similarly, the opposite can also be achieved by a reverse flow, for example in the medical domain where adversarial training was used to make real images more like synthetic images, and clinically-relevant features are preserved via self-regularization (Mahmood et al., 2018). The translated images were then interpreted by networks trained on large datasets of synthetic medical images.

The Cycle-GAN approach, applied in this paper, expands on this GAN approach to the domain of semantic segmentation, by allowing for unpaired image-to-image translation whilst re-using the synthetic dense pixel annotations. However, also this method is limited in ensuring the transfer of semantic information and has limited domain shift capability, for example when the classes between sets differ.

Recent results in parallel to this work have shown further progress to allow unsupervised domain adaptation for pixel-level semantic segmentation for improving the realism of synthetic datasets. Currently, novel methods are proposed that work both on the feature-level and pixel-level. The Cycle-Consistent Adversarial Domain Adaptation model (CyCADA) uses a cycle-consistency los similar to Cycle-GAN to ensure the preservation of pixel information but adds a semantic labeling loss to ensure semantic consistency (Hoffman et al., 2018). Applied to a synthetic-to-real scenario in a semantic segmentation task, the performance gap caused by the previous disparity in realism was almost closed.

Another effort in this direction is a domain adaptive semantic segmentation method, with the aim to learn a fully convolutional network semantic segmentation model which is adapted for use on the unlabelled target domain (Hoffman et al., 2016). The method uses domain adversarial learning for global domain alignment, while using a class-aware constrained loss for transferring the spatial layout.

The emergence of the field is further acknowledged by the recent release of benchmarks for synthetic-to-real visual domain adaptation (Peng et al., 2018).

### 1.3. Outline

The paper is structured in 2 parts, each with their corresponding materials, methods, discussion and conclusion sections. Part I describes the image-to-image translation from the synthetic rendered domain to the empirical photographic domain. Part II validates the proposed approach of Part I by studying the effect on segmentation learning, using the translated images from Part I. The paper ends with a general discussion and conclusion.

## 2. Part I: image-to-image translation

In this first part of the paper we describe and analyze the unpaired image-to-image translation on agricultural images from the synthetic to the empirical domain and vice versa. The main objective was to obtain pairs of images P consisting of an image translated from the synthetic to the empirical domain, and a corresponding ground truth map (see Fig. 1).

### 2.1. Materials

#### 2.1.1. Image dataset

The unpaired image dataset (Barth et al., 2017b) of *Capsicum annuum* (sweet- or bell pepper) was used that 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. In both sets, 8 classes were annotated on a per-pixel level, either manually for the empirical dataset or computed automatically for the synthetic dataset. In Fig. 2 examples of images in the dataset are shown. The dataset was publicly released at:

Both synthetic and empirical images were first cropped to 424x424

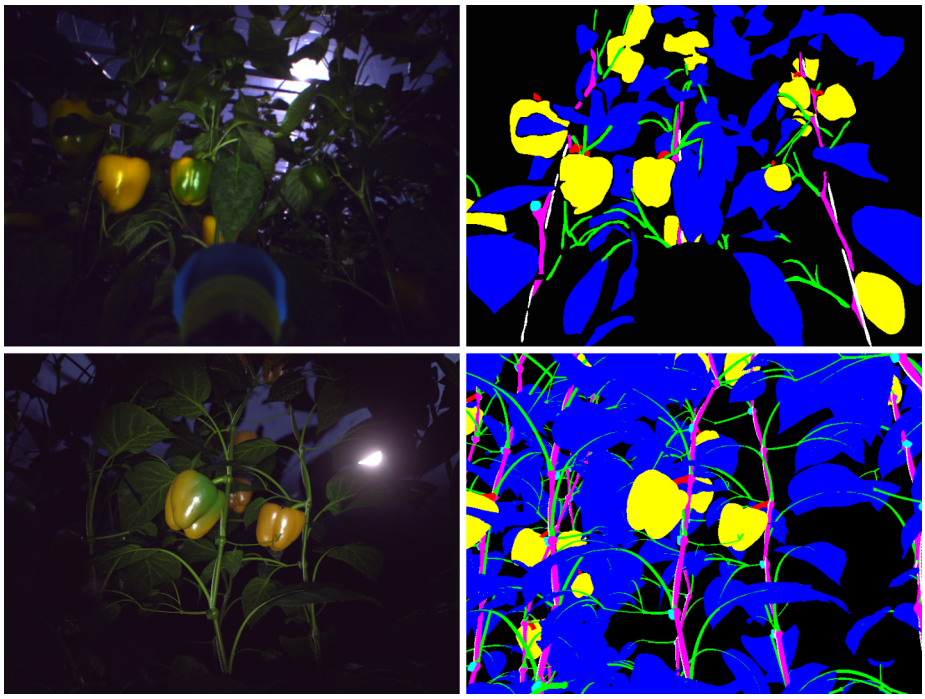

**Fig. 2.** Uncropped examples of empirical (top row) and synthetic (bottom row) color images (left column) and their corresponding ground truth labels (right column). Part class labels: ● background, ● leafs, ● peppers, ● peduncles, ● stems, ● shoots and leaf stems, ○ wires and ● cuts where pepper where harvested.

pixels to exclude the robot end-effector's suction cup in the image, because initial image-to-image translation experiments showed the cup was replicated undesirably in other parts of the image. This was in line with previous findings from the same authors where color and texture translation often succeeded, but domains with a large geometric variation were translated with less success (Zhu et al., 2017). Secondly, the image was resampled bilinearly to a resolution of 1000 × 1000 pixels as additional experiments during Part II showed upscaling improved the learning.

From the *Capsicum annuum* dataset, the synthetic images 1–1000 were used for training the model for translation, which was applied to translate the full dataset thereafter. For the empirical images, 50 annotated images of the *Capsicum annuum* dataset were used for testing, whereas for training 175 non-annotated images were used that were not part of the released dataset, but were collected during the same data acquisition experiment.

#### 2.1.2. Software

The Berkeley AI Research (BAIR) laboratory implementation of unpaired image-to-image translation using cycle-consistent adversarial networks was used (Zhu et al., 2017).

#### 2.1.3. 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.

### 2.2. Methods

The adversarial learning scheme in Fig. 1 was applied with synthetic images as domain X and empirical images as domain Y. The hyperparameters of the Cycle-GAN were manually optimised by visually evaluating the resulting images with their target domain. The number of generative and discriminative filters were set to 50 and the learning rate was set to 0.0002 with an ADAM (Kingma and Ba, 2017) momentum term of 0.5. The basic discriminator model was used, whereas for the generator the RESNET 6 blocks model was used (He et al., 2016). Weights for the cycle loss were set to 10 for each translation direction.

#### 2.2.1. Quantitative translation evaluation

Although the success of the translation is already quantitatively captured by the adversarial loss, this measure is biased and mathematically obfuscated. By specifically looking at key image features like color, contrast, homogeneity, energy and entropy, it could be derived if the translated images improved on those features. This would provide evidence about the dissimilarity gap between the synthetic and empirical domains.

For this purpose, we first compared for each object part class the synthetic color distribution prior and post translation with those of the empirical distribution. The color spectrum of each class was obtained by first transforming the color images to HSI colorspace. The Hue channel in the transformed image represented for each pixel which color was present, regardless of illumination and saturation intensity. The histogram of this channel was then taken to count the relative color occurrence per class.

As we hypothesise that the color difference between the synthetic and the empirical domain images will be reduced after translation of the synthetic images, the correlations of the color distributions of each object part class were compared for (i) the empirical images and the synthetic images, and (ii) the empirical images and the translated images.

Second, to obtain additional image features, first an average gray level co-occurrence matrix (GLCM) (Haralick et al., 1973) was calculated for each class for the first 10 images in the synthetic, synthetic translated to empirical and empirical sets. The GLCM summarises how often a pixel with a certain intensity value $i$ occurs in a specific spatial relationship to a pixel with the intensity value $j$. This relationship was set to address horizontally neighbouring pixels only. From the GLCM, the following features were derived:

**Contrast** $= \sum_{i,j} |i-j|^2 GLCM(i,j)$, measuring the overall difference in luminance between neighbouring pixels.
**Homogeneity** $= \sum_{i,j} \frac{GLCM(i,j)}{1+|i-j|}$, a value that measures the closeness of the distribution of elements in the GLCM to the GLCM diagonal, which implies that high values of homogeneity reflect the absence of changes in the image and indicates a locally homogenous distribution in image textures.
**Energy** $= \sum_{i,j} GLCM(i,j)^2$, a measure of texture crudeness or disorder.

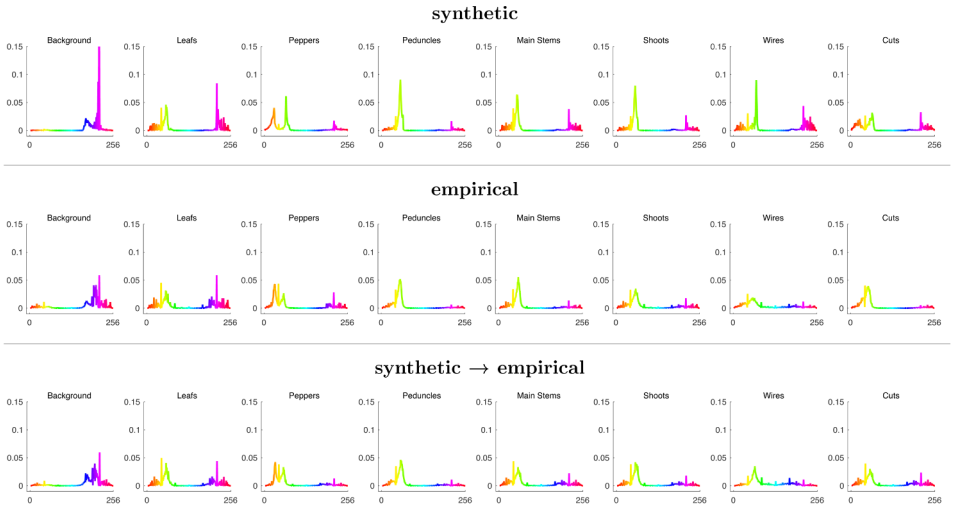


**Fig. 3.** Color distributions discretized to 256 values in the hue channel (x-axis) per class of the synthetic, empirical and synthetic translated to empirical images. Integral per distribution amounts to 1 (y-axis).

**Entropy** $= \sum_{i,j} -ln(GLCM(i,j)) \cdot GLCM(i,j)$, measuring the amount of information or complexity in the image.

### 2.3. Results

In Fig. 4 the results of the image-to-image translations are shown. The second column is of most interest to our research, as it shows the set $X_y$ of synthetic images which were translated to the empirical domain. However, as a reference also the translation from empirical to the synthetic domain is shown in the third column.

The color distributions for each object part class for the synthetic, empirical and translated synthetic images are shown in Fig. 3. The corresponding correlations between the empirical images and the synthetic as well as translated synthetic images are shown in Table 1.

For the image features contrast, homogeneity, energy and entropy, the results per class for the synthetic, empirical and synthetic translated to empirical images are shown in Fig. 5. The difference of 0.100 was found in contrast averaged over all classes between the synthetic and empirical set, whereas this difference was reduced to 0.015 for the translated and the empirical set. Similarly, for homogeneity this was reduced from 0.028 to 0.015. For the energy feature, this was reduced from 0.126 to 0.026. Regarding entropy, the average difference was reduced from 0.364 to 0.003.

### 2.4. Discussion and conclusion

Qualitative visual evaluation of the results in Fig. 4 showed a remarkable translation of synthetic images to empirical looking images and vice versa. Most notably the scattering of illumination and color of each plant part were converted realistically. It also appeared that the model learns to distinguish plant parts without any supervised information, as the (partially) ripe and unripe fruit were often translated to the other domain with altered maturity levels. A difference in camera focus seemed translated properly, indicating that local features (e.g. edge blur and texture) can be mapped accurately.

Some image artifacts did arise however, especially the translation of overexposed areas like sunshine or fruit reflections. The explanation might be that the model cannot generate this information correctly because any information beyond overexposure prior translation was already collapsed to a single maximum value (e.g. 255). Furthermore, an overlay of a checkerboard-like texture seems to have been added to the translated local textures, which might be preventable in the future (Odena et al., 2016).

The image-to-image translation method appeared not to be suited when one image set contained additional objects or parts that were absent in the other set, such as the presence of a suction cup in our earlier experiments. We noticed in previous experiments that this part was undesirably replicated in other areas of the image.

In Fig. 4 we can also see that large morphological features (e.g. large plant part shape and geometry relatively to other plant parts) were not translated, indicating a limitation of the Cycle-GAN approach. However, since geometry was not translated, this did allow for using the underlying synthetic ground truth labels to be used with the translated images for Part II. If also the geometry would have been translated, then the ground truth labels would not have been translated accordingly.

In Fig. 3, the translation effect on color distribution can be seen for each plant part and background. Quantitatively, the mean color correlation of 0.62 between the synthetic and empirical images increased post translation to 0.90 (see Table 1 for correlations per plant part and the mean over all plant parts). Indeed this is also what we observe in Fig. 4, where for example the color of the fruit in the translated synthetic images matches the empirical images more than those of the synthetic images.

When we look at the averages of the image texture features contrast, homogeneity, energy and entropy, they were closer together when comparing the empirical images and synthetic translated images than when comparing the empirical images and the synthetic images. For some individual classes this did not hold however, e.g. the homogeneity of the cuts was erroneously doubled instead. In Fig. 4 it can indeed be observed that local level textures of the translated synthetic images have become more similar to those of the empirical images. For

**Table 1**
Average color distribution correlations per object part class between (i) the empirical images and synthetic images, and (ii) the empirical images and the translated synthetic images.

| | backgr. | leafs | peppers | peduncles | stems | shoots | wires | cuts | mean |
|---|---|---|---|---|---|---|---|---|---|
| correlation(synthetic, empirical) | 0.25 | 0.78 | 0.42 | 0.93 | 0.76 | 0.83 | 0.45 | 0.48 | **0.62** |
| correlation(synthetic→empirical, empirical) | 0.86 | 0.94 | 0.93 | 0.93 | 0.92 | 0.98 | 0.81 | 0.79 | **0.90** |


**Fig. 4.** Image-to-image translation examples using Cycle-GAN. Source domain images prior translation are shown in the outer columns; synthetic images (left) and empirical images (right). The second column shows the set of interest $X_y$; the translated synthetic images to the empirical domain. The third column shows empirical images translated to synthetic domain.

example, the smoothness of the fruit in the translated synthetic images is improved towards the empirical images, as compared to the more coarse and grainy surface texture of the fruit in the synthetic images.

Regarding our first hypothesis, we therefore confirm that image feature differences with the empirical set were reduced after translation of the synthetic images, using a cycle-GAN.

This part of the work contributed to the field of computer vision (e.g. for agricultural robotics) by providing a method for optimising realism in synthetic training data to potentially improve state-of-the-art machine learning methods that semantically segment plant parts, as evaluated in Part II of this paper.

## 3. Part II: evaluation of translation on semantic segmentation

In order to evaluate and validate the results of the translated images, Part II evaluates the effect of using translated images on object part segmentation learning. Our second hypothesis states that by bootstrapping with translated images and empirical fine-tuning, the highest empirical performance can be achieved over methods that bootstrap with limited dataset size of (30) empirical images or a large set (8750) of synthetic images. With our third hypothesis in this paper, we reckon that without any empirical fine-tuning, learning can be improved with translated images as compared to using only synthetic bootstrapping.

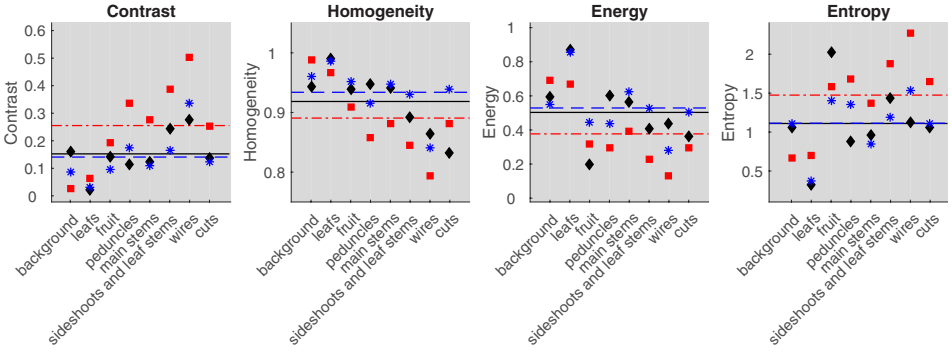


**Fig. 5.** Image features values for contrast, homogeneity, energy and entropy per class for the empirical ◆, synthetic ■ and synthetic translated to empirical ✳ images. Average over all classes is represented by a solid line for the empirical set, a dashed-dotted line for the synthetic set and a dashed line for the synthetic translated to empirical set.

### 3.1. Materials

The synthetic and empirical datasets as described in Part I (see Section 2.1.1) were used as well as the obtained image pairs $P(l_x, x_y)$ (see Fig. 1).

#### 3.1.1. Software

The publicly available semantic segmentation framework DeepLab V2 was used, which implemented convolutional neural network (CNN) models (Papandreou et al., 2015; Chen et al., 2015) on top of Caffe (Jia et al., 2014). Specifically, the VGG-16 network was used with a modification to include à *trous* spatial pyramid pooling for image context at multiple scales by convolutional feature layers with different fields-of-view (Chen et al., 2018; He et al., 2014).

#### 3.1.2. Hardware

Experiments were run on the same hardware as used in Part I. As a dependency for the DeepLab V2 Caffe version, the archived version of CUDA 7.5 was installed.

### 3.2. Methods

To compare performance differences, 7 experiments were performed using different combinations of train, fine-tune and test sets. The experiments were organised into 3 groups. The first group consisted of sanity checks and benchmarks, where in Experiments A, B and C were trained and tested with only a single type of data; empirical, synthetic and translated images respectively. The second group of experiments bootstrapped the CNN model with synthetic data and tested on empirical data, whilst either including (D) or omitting fine-tuning with empirical data (E). Group 3 consisted of similar experiments as Group 2, but interchanged the synthetic data with the domain translated images. The motivation for each experiment is given below and the used sets and image ranges for training, fine-tuning and testing are shown between brackets.

In order to contribute to the insights of what the effects are on the performance given the amount of training or fine-tuning images, all experiments are replicated whilst increasing the amount of data. To cope with the computational costs, only limited sample sizes could be explored. For the synthetic and translated data, the number of images per experiment have been increased each time with an order of magnitude ($10^1$, $10^2$, $10^3$, $10^{3.94} \approx 8750$). Given the small dataset size of the empirical set, the number of images were linearly increased (10, 20 and 30 samples). Below Experiments A through G are further specified

**Group 1: Sanity checks and benchmarks.**

**A** *Train Sets: empirical (#1–10 or #1–20 or #1–30).*
*Tes Set: empirical (#31–50).*
An experiment to see if the model can learn using only a small empirical dataset. This provides a reference for comparison of performance with other experiments that bootstrap with synthetic or translated synthetic images and/or fine-tune with empirical images. Given the small training size of the dataset in this experiment, the performance was expected to be low, compared to all other experiments that tested on empirical images.

**B** *Train Sets: synthetic (#1–10 or #1–100 or #1–1000 or #1–8750).*
*Test Set: synthetic (#8751–8800).*
This experiment was run to obtain baseline performance of the model when having access to a large and detailed annotated synthetic dataset. Performance is expected to be highest of all experiments because of the combination of perfect labels, large dataset size and relatively low image feature variance compared to empirical or synthetic translated images.

**C** *Train Sets: translated synthetic (#1–10 or #1–100 or #1–1000 or #1–8750).]*
*Test Set: translated synthetic (#8751–8800).*
This experiment was run to obtain baseline performance of the model when having access to a large and detailed annotated translated synthetic dataset. The performance should be similar of that of Experiment B, though is expected to a bit lower due to the extra variance that the empirical feature distribution might have introduced when synthetic images were translated to the empirical domain.

**Group 2: Synthetic bootstrap, empirical testing.**

**D** *Train Sets: synthetic (#1–10 or #1–100 or #1–1000 or #1–8750).*
*Test Set: empirical (#31–50).*
A reference experiment to see to what extent a network trained on synthetic images can generalise to the empirical domain, without fine-tuning with empirical images. Given the similarity gap between synthetic and empirical data, the performance should be relatively low compared to that of Experiment A or when compared to experiments that trained on a more realistic dataset, e.g translated synthetic as in Experiment F.

**E-1** *Train Sets synthetic (#1–10 or #1–100 or #1–1000 or #1–8750).*
*Fine-tune Set: empirical (#1–30).*
*Test Set: empirical (#31–50).*

**E-2** *Train Set: synthetic (#1–8750).*
*Fine-tune Sets: empirical (#1–10 or #1–20 or #1–30).*
*Test Set: empirical (#31–50).*
Similar to Experiment C, but with an extra fine-tuning step using empirical images. Performance is expected to be higher than C, because the network also optimises for the empirical image feature distribution. The performance of this experiment is expected to be lower than that of Experiment G, where the synthetic images were replaced by translated synthetic images, because the synthetic image feature distribution is more dissimilar with the empirical distribution than the translated synthetic distribution with the empirical distribution is.

**Group 3: Translated bootstrap, empirical testing.**

**F** *Train Sets: translated synthetic (#1–10 or #1–100 or #1–1000 or #1–8750).*
*Test Set: empirical (#31–50).*
With these experiments, we could check to what extent a synthetic trained network with improved realism can generalise to the empirical domain, without fine-tuning with empirical images. This experiment should provide the main result for our third hypothesis that states that without any fine-tuning with empirical images, improved learning for empirical images can be achieved using only translated images as opposed to using only synthetic images, as evaluated in Experiment C.

**G-1** *Train Sets: translated synthetic (#1–10 or #1–100 or #1–1000 or #1–8750).*
*Fine-tune Set: empirical (#1–30).*
*Test Set: empirical (#31–50).*

**G-2** *Train Set: translated synthetic (#1–8750).*
*Fine-tune Sets: empirical (#1–10 or #1–20 or #1–30).*
*Test Set: empirical (#31–50).*
These experiments should provide the main result for our second hypothesis, that states the synthetic images translated to the empirical domain can be used for improved learning of empirical images, as compared to using only synthetic images for bootstrapping (Experiment D). Performance of this experiment was expected to be the highest amongst all our experiments that tested on empirical data, because a large dataset with high similarity with the empirical images was used in combination with fine-tuning on empirical images.

#### 3.2.1. CNN training

For each experiment, a convolutional neural network was trained and/or fine-tuned and tested according to the dataset scheme as described in Section 3.2.

The hyperparameters of the network were manually optimised using separate validation datasets for combination of models and data set configurations as suggested by Goodfellow et al. (2016) and Bengio (2012). This resulted in using Adaptive Moment Estimation (ADAM) (Kingma and Ba, 2017) 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 previously (Barth et al., 2017a) and therefore we fixed them across conditions. An adjustment was made in the layer weight initialisation procedure, by updating the model to using MSRA weight fillers (He et al., 2015; Mishkin and Matas, 2019). Furthermore, the dropout rate (Srivastava et al., 2014) was adjusted to 0.50 to circumvent early overfitting and facilitate generalisation. The size of the input layer was cropped to 929x929 pixels, which was the maximum that our GPU memory could handle.

#### 3.2.2. Performance evaluation

To calculate the performance of the segmentation, we used the Jaccard Index similarity coefficient as an evaluation procedure, also known as the intersection-over-union (IOU) (He and Garcia, 2009) which is widely used for semantic segmentation evaluation (Gabriela Csurka and Larlus, 2013; Everingham et al., 2010).

### 3.3. Results

In Fig. 8 the average IOU over all classes for Experiments A through G is shown. The results regarding the relationship between dataset size and performance for each experiments is shown in Fig. 7. In Fig. 6 the performances were split over the object part classes. Qualitative results are presented in Fig. 9.

### 3.4. Discussion and conclusion

#### 3.4.1. Group 1

In Experiment A, the aim was to investigate how the model learns to segment empirical images using only a small empirical training dataset. This provided a benchmark result and was expected to provide a lower bound of the learning for experiments where empirical learning or fine-tuning was used. Indeed compared to Experiments E and G the performance was lower. Furthermore, the performance of Experiment A was expected to be higher than experiments that tested on empirical data but did not use any during training or finetuning. Indeed this is confirmed by the lower performance of Experiments D and F.

Regarding the effect of dataset size, we expected an increase of performance with more empirical training samples. This is confirmed by the results in Fig. 7. Furthermore it can be noted that the increase in performance did not level out after 30 samples, suggesting that more empirical data would likely increase the performance further. The latter was anticipated, given that 30 images is a relatively small amount to capture all the variance that occurs in the context of our use-case.

Looking at the per class performance distribution of Experiment A in Fig. 6, it can be noted that the *cut* class was barely recognised having an IOU of 0.04. Recognising all classes was previously considered as a requirement (Barth et al., 2017a). Therefore it can be concluded that training with a small set of empirical images alone did not suffice, although qualitative results (see Fig. 9) looked promising and useful for some tasks like fruit detection.

In Experiment B, a baseline performance of the model was obtained when training and testing with the same large and detailed annotated dataset. Performance was expected to be highest of all experiments because of the perfect labels, large dataset size and because no domain shift occurred. Furthermore, the synthetic dataset has a relatively low image and geometric feature variance compared to empirical or synthetic translated images, which was likely to increase the learnability of the dataset. Indeed B achieved the best performance with an average IOU of 0.64. This performance could be used as a baseline to indicate the maximum obtainable IOU for this domain and currently used CNN architecture. Qualitative results still showed some gaps in thin and elongated classes like leaf stems and shoots, although results were much improved over previous segmentations where such gaps were larger (Barth et al., 2017a).

Concerning the effect of dataset size on performance, results in

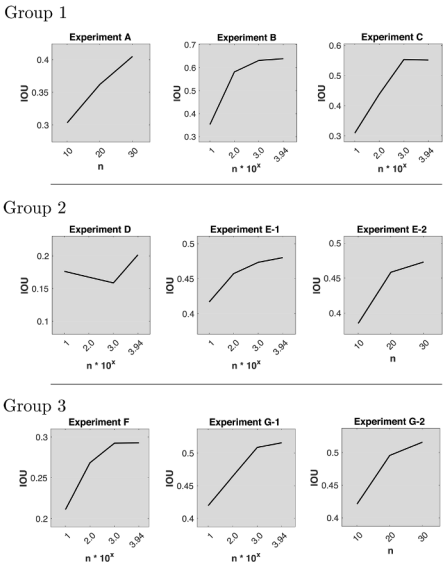


**Fig. 7.** Average IOU performance over test set for Experiments A through G-2, with increasing amounts of training or fine-tuning data (n).

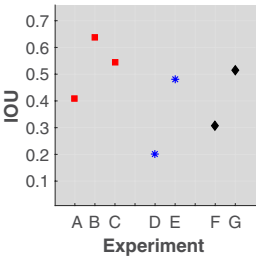


**Fig. 8.** Maximum average IOU performance over test set for Experiments A through G for Groups 1■, 2✳ and 3◆.

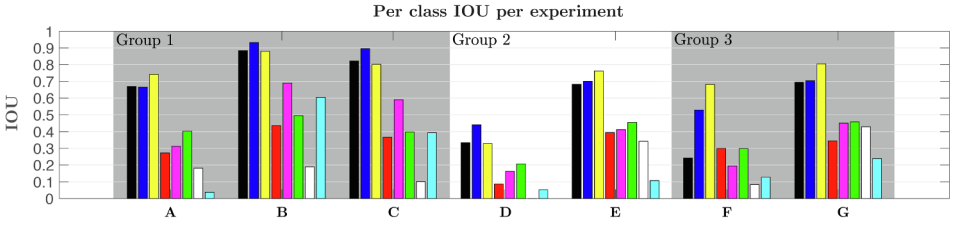


**Fig. 6.** For Experiments A through G, the IOU per class is displayed, ordered as: ● background, ● leafs, ● peppers, ● peduncles, ● stems, ● shoots and leaf stems, ○ wires and ● cuts where pepper where harvested.

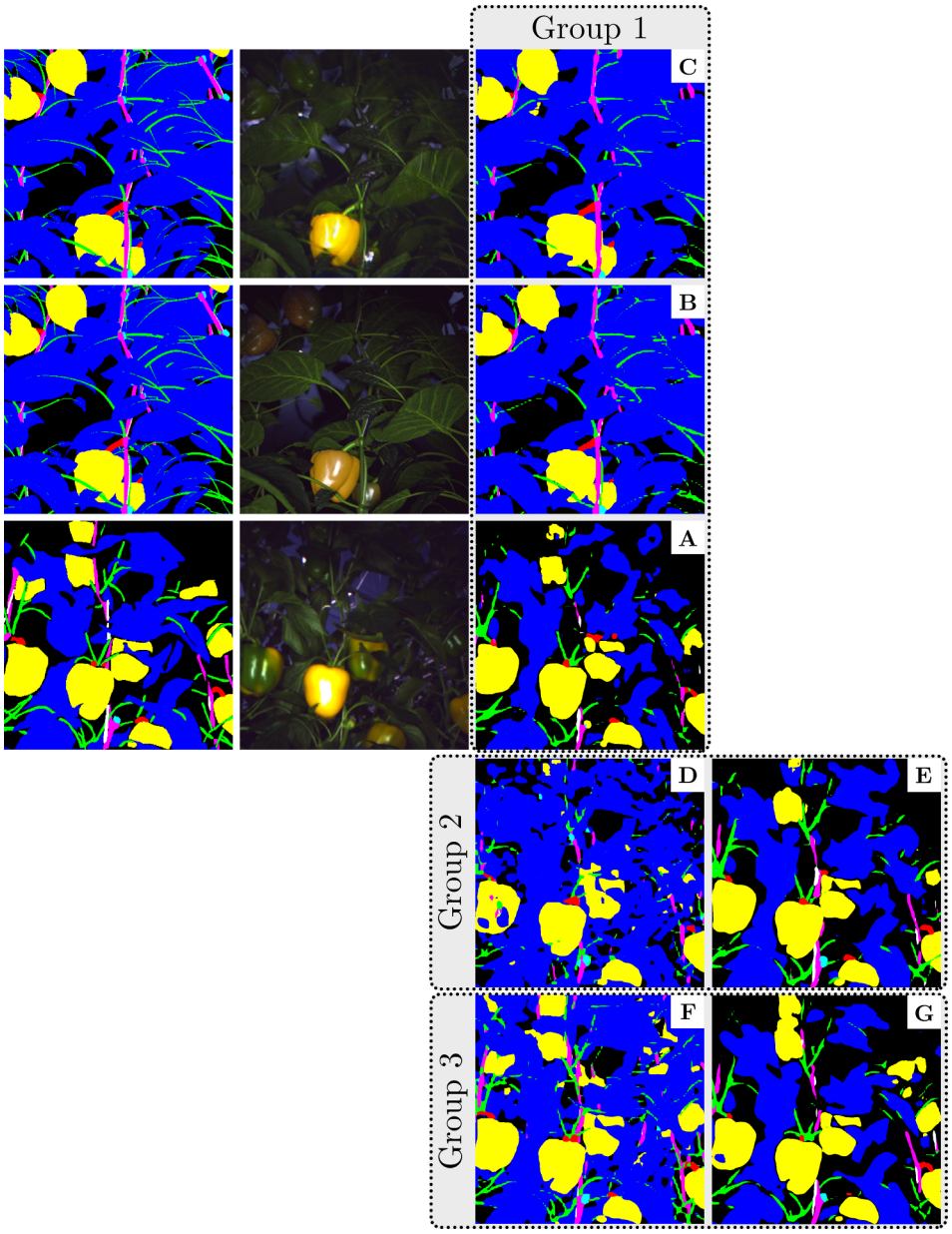


**Fig. 9.** Qualitative results from Experiments A through G. In the first column, the ground truths for synthetic translated to empirical (top), synthetic (middle) and empirical (bottom) are shown with labels: ● background, ● leafs, ● peppers, ● peduncles, ● stems, ● shoots and leaf stems, ○ wires and ● cuts. In the second column, their respective color images are displayed. Experimental results with models trained on the maximum amount of images, are grouped in the third and fourth column.

Fig. 7 show that the model stabilised in performance around a thousand images.

Experiment C was the last experiment from Group 1 which similar to B trained and tested with a large dataset, but instead used the translated images. The average IOU was expected to be relatively high, but lower than Experiment B due to the extra variance and noise that the empirical feature distribution introduces when synthetic images were translated to the empirical domain. Indeed the IOU was lower with 0.55 in C than the 0.64 in B, although the qualitatively results looked comparable. The effect of dataset size on performance was similar to B.

### 3.4.2. Group 2

The second group of experiments bootstrapped the CNN model with synthetic data and tested on empirical data, whilst either including (D) or omitting fine-tuning with empirical data (E).

Experiment D was a reference experiment to see to what extent a network trained on synthetic images can generalise to the empirical domain, without fine-tuning with empirical images. With an average IOU of 0.20, the performance approximately doubled over previous results where 424 × 424 pixel images were used (Barth et al., 2017a). However, the performance was lower than that of Experiment A, as expected because the synthetic data was relatively dissimilar with the empirical images.

Looking at the per class distribution, classes like *peduncle* and *cut* were barely recognised (IOU<0.1) and the *wire* class was omitted all together. Qualitatively, results looked far from similar to the ground truth and it can be concluded that training only with synthetic data would not be sufficient for many tasks, given the current learning architecture. There seems little relation between dataset size and performance. It could be that the domain gap is too large; adding more synthetic data does not help much in bridging the gap directly.

In Experiment E an extra fine-tuning step using empirical images was included. The IOU performance on empirical images was increased to 0.48 compared to Experiment A where no bootstrapping was applied; a 17% increase. Furthermore, looking at Fig. 6, it can be noted that recognition of all classes were now included, although the *cut* class was barely recognised (IOU = 0.11). Qualitatively, results looked close to the ground truth. Regarding the effect of dataset size, the performance increases but seems to level off at 8750 synthetic images (E-1) but is

expected to slightly further increase after 30 empirical images (E-2). Hence, fine-tuning seems to benefit from increased data of both types.

#### 3.4.3. Group 3

With the third group, similar experiments as Group 2 were performed, but interchanged the synthetic data with the domain translated images. Again a distinction was made with (F) and without (G) fine-tuning on empirical images.

Experiment F showed to what extent a synthetic trained network with improved realism can generalise to the empirical domain, without yet fine-tuning with empirical images. Compared to Experiment D (using synthetic images instead of translated ones), the performance increased with 55% to an average IOU of 0.31. This experiment confirms our third hypothesis that without any fine-tuning with empirical images, improved learning for empirical images can be achieved using only translated images as opposed to using only synthetic images. Although qualitatively, also improvements could be observed over Experiment D, it can be noted from the class performance distribution there existed still a relative poor performance on the classes *wires* and *cuts*.

In Experiment G, the model from Experiment F was fine-tuned with empirical images. This experiment should provide the main result for our second hypothesis, that states that synthetic images translated to the empirical domain can be used for improved learning of empirical images, as compared to using only synthetic images for bootstrapping, as evaluated in Experiment E. Our hypothesis is confirmed by achieving the best performance on empirical data of an IOU = 0.52. This was an increase of 27% over Experiment A (only training on empirical images) and 8% improvement over Experiment E.

Qualitatively, results in Experiment G looked comparable to results of Experiments A and E, but more comparable to the ground truth. Looking at the class distribution, all classes were included. Most notably the *cut* class performance increased with 118% over Experiment E and with 600% over Experiment A to an IOU of 0.24.

To summarise, without using any annotated empirical training images, an improved performance can be achieved by bootstrapping with translated synthetic images (F) compared to the model that only uses synthetic images (D). Moreover, we have shown that it ensures improved recognition of minor object part classes like *wires* and *cuts*. Additionally, we have shown that including fine-tuning with a small empirical dataset, with a model that has been bootstrapped on translated data, the highest performance on empirical images can be achieved (G).

## 4. General discussion and conclusion

In Part I, a cycle consistent generative adversarial network was applied to synthetic and empirical images with the objective to generate more realistic synthetic images by translating them to the empirical domain. Our analysis showed that the image feature distributions of these translated images, both in color and texture, were improved towards the empirical images. Regarding our first hypothesis, it was confirmed that the image feature difference with the empirical set was reduced after translation of the synthetic images. Qualitatively, the translated synthetic images looked highly similar to the real world images. However, some translation artifacts appeared. Furthermore the Cycle-GAN method could not improve upon geometric dissimilarities between the synthetic and the empirical domain. The latter proved an advantage however, as the synthetic ground truth also corresponded to the translated color images, allowing for the experiments on improved learning in the second part of our work.

In Part II, it was evaluated to what extent translated synthetic images to the empirical domain could improve on CNN learning with empirical images over other learning strategies. We confirmed our second hypotheses that by using translated images and fine-tuning with empirical images, the highest performance for empirical images can be achieved (IOU = 0.52) compared to training with only empirical (IOU = 0.41) or synthetic data (IOU = 0.48)

Besides improving segmentation performance on empirical images using translated synthetic images instead of only empirical or synthetic images during training, another key contribution of our work is the further minimisation of the CNN's dependency on annotated empirical data. We confirmed our third hypothesis that without any empirical image fine-tuning, learning can be improved with translated images (IOU = 0.31), a 55% increase over just using synthetic images (IOU = 0.20).

The work presented in this paper can be seen as an important step towards improved computer vision for domains such as agricultural robotics, medical support systems or autonomous navigation. It facilitates CNN semantic object part segmentation learning without or minimal requirement of manually annotated images and ensures improved recognition of minor parts.

## Credit authorship contribution statement

**R. Barth:** Conceptualization, Methodology, Software, Formal analysis, Data curation, Funding acquisition, Writing - original draft, Visualization. **J. Hemming:** Validation, Writing - review & editing, Funding acquisition, Project administration. **E.J. Van Henten:** Validation, Writing - review & editing, Supervision.

## Acknowledgements

This research was partially funded by the European Commission in the Horizon2020 Programme (SWEEPER GA No. 644313) and the Dutch Ministry of Economic Affairs (EU140935).

## References

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. https://doi.org/10.1016/j.compag.2013.05.004. http://www.sciencedirect.com/science/article/pii/S0168169913001099.

Bac, C.W., van Henten, E.J., Hemming, J., Edan, Y., 2014. Harvesting robots for high-value crops: state-of-the-art review and challenges ahead. J. Field Robot. 31 (6), 888–911. https://doi.org/10.1002/rob.21525.

Barth, R., IJsselmuiden, J., Hemming, J., Henten, E.V., 2017. Synthetic bootstrapping of convolutional neural networks for semantic plant part segmentation, Comput. Electron. Agric. doi:https://doi.org/10.1016/j.compag.2017.11.040.

Barth, R., IJsselmuiden, J., Hemming, J., Henten, E.V., 2017. Data synthesis methods for semantic segmentation in agriculture: a capsicum annuum dataset, Comput. Electron. Agric. doi: 10.1016/j.compag.2017.12.001.

Bengio, Y., 2012. Practical Recommendations for Gradient-Based Training of Deep Architectures. Springer, Berlin, Heidelberg, pp. 437–478. https://doi.org/10.1007/978-3-642-35289-8_26.

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., Schroff, F., Adam, H., 2017. Rethinking Atrous Convolution for Semantic Image Segmentation. ArXiv e-printsarXiv: 1706.05587.

Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2018. Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. 40 (4), 834–848. https://doi.org/10.1109/TPAMI.2017.2699184.

Csurka, G., 2017. A Comprehensive Survey on Domain Adaptation for Visual Applications. Springer International Publishing, Cham, pp. 1–35. https://doi.org/10.1007/978-3-319-58347-1_1.

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

Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A., 2010. The pascal visual object classes (voc) challenge. Int. J. Comput. Vision 88 (2), 303–338.

Gabriela Csurka, F.P., Larlus, Diane, 2013. What is a good evaluation measure for semantic segmentation? In: Proceedings of the British Machine Vision Conference. BMVA Press.

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

papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.
Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press. http://www.deeplearningbook.org.
Haralick, R.M., Shanmugam, K., Dinstein, I., 1973. Textural features for image classification. IEEE Trans. Syst. Man Cybernet. SMC-3 (6), 610–621. https://doi.org/10.1109/TSMC.1973.4309314.
He, H., Garcia, E.A., 2009. Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. 21 (9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239.
He, K., Zhang, X., Ren, S., Sun, J., 2014. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. Springer International Publishing, Cham, pp. 346–361. https://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. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), ICCV '15. IEEE Computer Society, Washington, DC, USA, 2015, pp. 1026–1034. doi:10.1109/ICCV.2015.123.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.
Hoffman, J., Wang, D., Yu, F., Darrell, T., 2016. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. CoRR abs/1612.02649. arXiv:1612.02649. http://arxiv.org/abs/1612.02649.
Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., Efros, A., Darrell, T., 2018. CyCADA: Cycle-consistent adversarial domain adaptation. In: Dy, J., Krause, A. (Eds.), Proceedings of the 35th International Conference on Machine Learning, vol. 80 of Proceedings of Machine. Learning Research, PMLR, Stockholmsmssan, Stockholm Sweden, pp. 1989–1998 http://proceedings.mlr.press/v80/hoffman18a.html.
Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5967–5976.
Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T., 2014. Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the 22Nd ACM International Conference on Multimedia, MM '14. ACM, New York, NY, USA, 2014, pp. 675–678. doi:10.1145/2647868.2654889. URL http://doi.acm.org/10.1145/2647868.2654889.
Kingma, D.P., Ba, J., 2017. Adam: A method for stochastic optimization. CoRR abs/1412.6980. http://arxiv.org/abs/1412.6980.
Long, M., Cao, Y., Wang, J., Jordan, M.I., 2015a. Learning transferable features with deep adaptation networks. In: Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37, ICML'15. JMLR.org, pp. 97–105 http://dl.acm.org/citation.cfm?id=3045118.3045130.
Long, J., Shelhamer, E., Darrell, T., 2015b. Fully convolutional networks for semantic segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
Mahmood, F., Chen, R., Durr, N.J., 2018. Unsupervised reverse domain adaptation for synthetic medical images via adversarial training. IEEE Trans. Med. Imaging 1. https://doi.org/10.1109/TMI.2018.2842767.
Mishkin, D., Matas, J., 2019. All you need is a good init. CoRR abs/1511.06422. http://arxiv.org/abs/1511.06422.
Odena, A., Dumoulin, V., Olah, C., 2016. Deconvolution and checkerboard artifacts, Distill. http://distill.pub/2016/deconv-checkerboard/.
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.
Peng, X., Usman, B., Saito, K., Kaushik, N., Hoffman, J., Saenko, K., 2018. Syn2real: A new benchmark forsynthetic-to-real visual domain adaptation. CoRR abs/1806.09755.
Qiu, W., Yuille, A., 2016. Unrealcv: connecting computer vision to unreal engine. In: Hua, G., Jégou, H. (Eds.), Computer Vision – ECCV 2016 Workshops. Springer International Publishing, Cham, pp. 909–916.
Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., Lawrence, N.D., 2019. Dataset shift in machine learning.
Richter, S.R., Vineet, V., Roth, S., Koltun, V., 2016a. Playing for data: ground truth from computer games. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (Eds.), European Conference on Computer Vision (ECCV). LNCS, vol. 9906. Springer International Publishing, pp. 102–118.
Richter, S.R., Vineet, V., Roth, S., Koltun, V., 2016b. Playing for data: ground truth from computer games. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (Eds.), European Conference on Computer Vision (ECCV). LNCS, vol. 9906. Springer International Publishing, pp. 102–118.
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.
Saenko, K., Kulis, B., Fritz, M., Darrell, T., 2010. Adapting visual category models to new domains. In: Daniilidis, K., Maragos, P., Paragios, N. (Eds.), Computer Vision – ECCV 2010. Springer, Berlin, Heidelberg, pp. 213–226.
Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A., 2011. Real-time human pose recognition in parts from single depth images. In: CVPR 2011, pp. 1297–1304. doi: 10.1109/CVPR.2011.5995316.
Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., Webb, R., 2017. Learning from simulated and unsupervised images through adversarial training. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2242–2251. https://doi.org/10.1109/CVPR.2017.241.
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.
Sun, B., Saenko, K., 2016. Deep coral: Correlation alignment for deep domain adaptation. In: Hua, G., Jégou, H. (Eds.), Computer Vision – ECCV 2016 Workshops. Springer International Publishing, Cham, pp. 443–450.
Wood, E., Baltrusaitis, T., Morency, L.-P., Robinson, P., Bulling, A., 2016. Learning an appearance-based gaze estimator from one million synthesised images. In: Proc. of the 9th ACM International Symposium on Eye Tracking Research & Applications (ETRA 2016), 2016, pp. 131–138. doi:10.1145/2857491.2857492. https://perceptual.mpi-inf.mpg.de/files/2016/01/wood16_etra.pdf.
Zhu, J.-Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2242–2251.