Proceedings of the IEEE IROS workshop on

# Agricultural Robotics

## learning from Industry 4.0 and moving into the future

September 28, 2017

Vancouver, Canada

Tsampikos Kounalakis
Frits van Evert
David Michael Ball
Gert Kootstra
Lazaros Nalpantidis

https://doi.org/10.18174/434834

# Optimising Realism of Synthetic Agricultural Images using Cycle Generative Adversarial Networks.

Ruud Barth[1], Joris IJsselmuiden[2], Jochen Hemming[1] and Eldert J. van Henten[2]

***Abstract*— A bottleneck of state-of-the-art machine learning methods, e.g. deep learning, for plant part image segmentation in agricultural robotics is the requirement of large manually annotated datasets. As a solution, large synthetic datasets including ground truth can be rendered that realistically reflect the empirical situation. However, a dissimilarity gap can remain between synthetic and empirical data by incomplete manual modelling. This paper contributes to closing this gap by optimising the realism of synthetic agricultural images using unsupervised cycle generative adversarial networks, enabling unpaired image-to-image translation from the synthetic to empirical domain and vice versa. For this purpose, the Capsicum annuum (sweet- or bell pepper) dataset was used, containing 10,500 synthetic and 50 empirical annotated images. Additionally, 225 unlabelled empirical images were used. We hypothesised that the similarity of the synthetic images with the empirical images increases qualitatively and quantitively when translated to the empirical domain and investigated the effect of the translation on the factors color, local texture and morphology. Results showed an increased mean class color distribution correlation with the empirical dataset from 0.62 prior and 0.90 post translation of the synthetic dataset. Qualitatively, synthetic images translate very well in local features such as color, illumination scattering and texture. However, global features like plant morphology appeared not to be translatable.**

## I. INTRODUCTION

A key success factor of agricultural robotics performance is a robust underlying perception methodology that can distinguish and localise object parts [1], [2], [3]. In order to train state-of-the-art machine learning methods that can achieve this feat, large empirical annotated datasets are required. Synthetic data can help bootstrapping such methods in order to reduce the required amount of empirical data [4]. However, a gap in realism remains between the modelled synthetic data and the empirical images, plausibly restraining synthetic bootstrapping performance.

In this paper we report on optimising the realism of synthetic images modelled from empirical data [5]. The long term goal of this research is to improve plant part segmentation performance by synthetically bootstrapped deep convolutional neural networks (CNN) [4]. For the intermediate goal presented here, we hypothesise the dissimilarity between corresponding synthetic and empirical images can be qualitatively and quantitatively reduced using unpaired image-to-image translation by cycle-consistent adversarial networks (cGAN) [6].

Convolutional neural networks currently show state-of-the-art performance on image segmentation tasks [7], [8], [9]. However, CNNs require large annotated datasets on a per-pixel level in order to successfully train 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 dataset size dependency. Satisfying this requirement can quickly become a bottleneck for learning.

One solution is to bootstrap CNNs with synthetically generated images including automatically computed ground truth [10], [11]. Consequently the bootstrapped network can be fine-tuned with and applied to empirical images, resulting in increased performance over methods without synthetic bootstrapping [4].

Previously we have shown methods to create such a dataset by realistically rendering 3D modelled plants [5]. Despite intensive manual optimisation for geometry, color and textures, we have shown that a discrepancy remains between synthetic and empirical data. Although this dataset can be used for successful synthetic bootstrapping and learning empirical images, there remains a gap between achieved performance and theoretical optimal performance [4].

Recently, the advent of generative adversarial networks (GAN) introduced another method of image data generation [12]. In GANs two models 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 of both models 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.

In later approaches, GANs were also conditioned by additional input images; both the generator and discriminator observe an input image [13]. The discriminator's goal is to compare such pairs on coherency of their co-occurrence whereas the generator aims to create an image-to-image translation from the conditional image to an image adhering

[1]Ruud Barth and Jochen Hemming are with Wageningen University & Research, Greenhouse Horticulture, P.O. Box 644, 6700 AP, Wageningen, The Netherlands. ruud.barth@wur.nl and jochen.hemming@wur.nl

[2]Joris IJsselmuiden and Eldert J. van Henten are with Wageningen University & Research, Farm Technology Group, Droevendaalsesteeg 1, 6708 PB, Wageningen, The Netherlands. joris.ijsselmuiden@wur.nl and eldert.vanhenten@wur.nl

to the same coherency of the other pairs in the dataset. The result is a generator $G$ that can translate images from one domain X (e.g. summer photographs) to images in another Y (e.g. winter photographs), formally notated as $G : X \rightarrow Y$.

A requirement for conditional GANs is the availability of geometrically paired images, but for many tasks these will not be available. For example in agriculture, obtaining a geometrically paired synthetic image of an empirical scene would defeat the purpose of circumventing manual annotation time. Instead, only unpaired synthetic images can be generated without additional manual efforts.

A recent approach aimed to dissolve this requirement by investigating unpaired image-to-image translation. In cycle-consistent adversarial networks (Cycle-GAN) [6], a mapping $G : X \rightarrow Y$ is learned whilst also an inverse mapping $F : Y \rightarrow X$. 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 $D_Y$ safeguards a 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 an geometrically similar image in domain $Y$ (or $y \in Y$ to $X$). This 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 the original geometry is maintained.

To achieve that, a cycle consistency loss was added to further regularise the learning. 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 image is translated from one domain to the other and back again, an equal image should be retrieved. This forces the generators $G$ and $F$ to achieve unpaired geometrically consistent image-to-image translation from one domain to the other and vice versa.

The key contribution of our research presented here is that we show that Cycle-GAN can translate agricultural images in the synthetic domain to images in the empirical domain, to improve the realism of the synthetic data and close the dissimilarity gap further. Hence, this will increase the amount of realistic training data for machine learning computer vision methods. This can be seen as an important step towards improved sensing for agricultural robotics by minimising the dependency on manual annotated datasets.

The scope of this paper was limited to results of the translation and a similarity comparison, whereas future research will investigate the impact of translated synthetic images on learning.

## II. MATERIALS

### A. Image dataset

The unpaired image dataset of *Capsicum annuum* (sweet- or bell pepper) was used [5] 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 and geometrically. In both sets 8 classes were annotated on a per-pixel level,

Fig. 1: Uncropped examples of empirical and synthetic color images (2nd and 3rd image respectively) and their corresponding ground truth labels. Class labels: ● background, ● leafs, ● peppers, ● peduncles, ● stems, ● shoots and leaf stems, ○ wires and ● cuts where pepper where harvested.

either manually for the empirical dataset or automatically for the synthetic dataset. In Figure 1 examples of images in the dataset are shown. The dataset was publicly released at: `http://dx.doi.org/10.4121/uuid:884958f5-b868-46e1-b3d8-a0b5d91b02c0`

Both synthetic and empirical images were cropped to 424x424 pixels to exclude the robot end-effector's suction cup in the image, because initial image-to-image translation experiments showed that this hardware was replicated undesirably in other parts of the image. This is in line with comments of the methodology of the Cycle-GAN authors; color and texture translation often succeeds though large geometric changes are translated with less success.

From the *Capsicum annuum* dataset, the synthetic images 1-1,000 were used for training and the remainder for testing. For the empirical images, an unreleased and unlabelled dataset consisting of 225 images was used, of which a random subset was previously labelled and included in the *Capsicum annuum* dataset. 175 Images of this set were used for training and 50 for testing.

### B. Software

The Berkeley AI Research (BAIR) laboratory implementation of unpaired image-to-image translation using cycle-consistent adversarial networks was used [6].

## III. METHODS

### A. Image-to-Image Translation

Hyper-parameters of the Cycle-GAN were manually optimised. The full resolution of the cropped images was used. The number of generative and discriminative filters were set to 50 and the learning rate was set to 0.0002 with an ADAM [14] momentum term of 0.5. The basic discriminator model was used, whereas for the generative model the RESNET [15] 6 blocks model. Weights for the cycle loss were set to 10 for each direction.

### B. Quantitative Translation Comparision

Although the success of the translation will already be quantitatively captured by the adversarial loss, this measure is biased and mathematically obfuscated. It is interesting to look more specifically at key image features like color.

For this purpose, we compared for each class the synthetic color distribution prior and post translation with those of the empirical distribution. We hypothesise the color difference post translation will be reduced. To determine this quantitatively, the correlation between color distributions was obtained.

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, irregardless of illumination and saturation intensity. The histogram of this channel was then taken to count the relative color occurrence per class.

## IV. RESULTS

In Figure 3 the results of the image-to-image translations are shown. The second column is of most interest to our research, as it shows synthetic images which are translated to the empirical domain. However, as reference also the

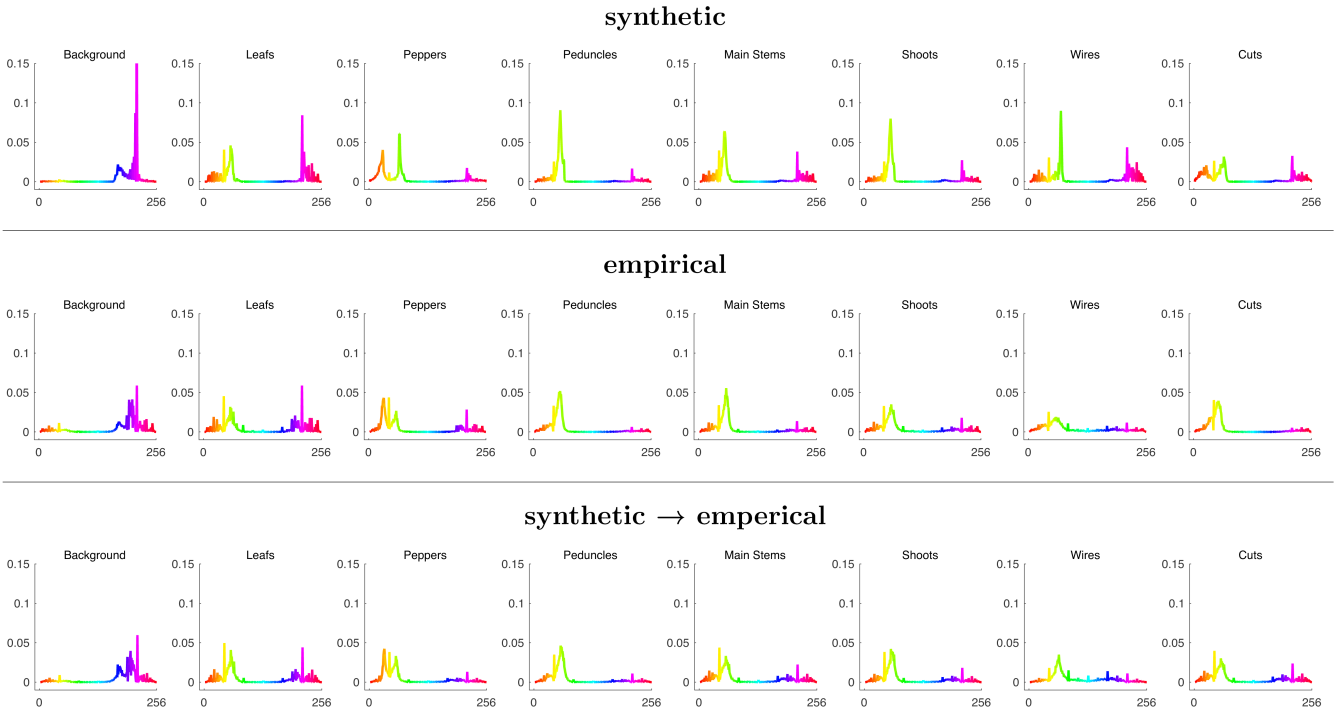


Fig. 2: Color distributions (discretized to 256 values in the hue channel) per class of the synthetic, empirical and synthetic translated to empirical images. Integral per distribution amounts to 1.

| | backgr. | leafs | peppers | peduncles | stems | shoots | wires | cuts | avg. |
|---|---|---|---|---|---|---|---|---|---|
| **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** |

TABLE I: Color distribution correlation per class between the synthetic and synthetic translated to empirical with the empirical image dataset.


Fig. 3: Image-to-image translation examples using Cycle-GAN. Source images prior translation are shown in the outer columns; synthetic images (left) and empirical images (right). The second column shows translated synthetic images to empirical ones and the third column shows empirical images translated to synthetic ones.

translation from empirical to the synthetic domain is shown in the third column.

The color distributions for each class for the synthetic, empirical and synthetic→empirical translation are shown in Figure 2. The corresponding correlations between the empirical images and the synthetic or synthetic→empirical images are shown in Table I.

## V. DISCUSSION

Qualitative evaluation of the results subjectively showed a remarkable feat of translated synthetic images to empirical looking images and vice versa. Notably the scattering of illumination and color of each plant part were converted realistically. It appears the model learns to distinguish plant parts without any supervised information, as often the (partially) ripe and unripe fruit are 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.

Artifacts do arise however, especially the translation to overexposed area's like sunshine or fruit reflections. The explanation might be that the model cannot generate this information correctly because information beyond the maximum range of the image was previously collapsed into a single the maximum value (e.g. 255) of the image. Furthermore, a faint checker-like texture seems to have been added to the translated local textures.

Larger morphological features (e.g. plant part shape and their geometry) were not translated, indicating a limitation of the Cycle-GAN approach. This suggests that the source synthetic data should be geometrically highly similar to the empirical situation, for a realistic translation to succeed.

If the translated images are later to be used for supervised learning, the morphological structure should be retained however. This because the underlying ground truth cannot be translated correspondingly, as no supervision is used in Cylce-GAN.

The method is not suited when one image set contains additional parts absent in the other set, e.g. the inclusion of a suction cup in our earlier experiments.

In Figure 2 the translation effect on color distribution can be seen for each plant part and background. Quantitatively, the mean correlation of the color distributions increased of the synthetic data with the empirical data prior (0.62) and post translation (0.90), confirming our hypothesis that color difference post translation with the empirical data is reduced.

## VI. CONCLUSION

This work contributed to the field of agricultural robotics by providing a method for optimising realism in synthetic training data to improve state-of-the-art machine learning methods that semantically segment plant parts.

Our hypothesis that dissimilarity between synthetic and empirical images can be reduced by using adversarial generative networks (e.g. Cycle-GAN) has been confirmed qualitatively and quantitatively by increasing the color distribution correlation with empirical images prior and post translation of synthetic images.

Future research will investigate the impact on learning with empirically translated synthetic images. Due to the improved realism, it might become feasible to circumvent the need of any manual annotation of empirical data by solely bootstrapping on translated synthetic data, without the requiring empirical fine-tuning.

## ACKNOWLEDGMENT

This research was partially funded by the European Commission in the Horizon2020 Programme (SWEEPER GA no. 644313).

## REFERENCES

[1] C. Bac, J. Hemming, and E. van Henten, "Robust pixel-based classification of obstacles for robotic harvesting of sweet-pepper," *Computers and Electronics in Agriculture*, vol. 96, pp. 148 – 162, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0168169913001099

[2] A. Gongal, S. Amatya, M. Karkee, Q. Zhang, and K. Lewis, "Sensors and systems for fruit detection and localization: A review," *Computers and Electronics in Agriculture*, vol. 116, pp. 8 – 19, 2015.

[3] C. W. Bac, E. J. van Henten, J. Hemming, and Y. Edan, "Harvesting robots for high-value crops: State-of-the-art review and challenges ahead," *Journal of Field Robotics*, vol. 31, no. 6, pp. 888–911, 2014. [Online]. Available: http://dx.doi.org/10.1002/rob.21525

[4] R. Barth, J. IJsselmuiden, J. Hemming, and E. van Henten, "Synthetic bootstrapping of convolutional neural networks for semantic plant part segmentation," 2017, submitted to the Journal of Computers and Electronics in Agriculture.

[5] R. Barth, J. IJsselmuiden, J. Hemming, and E. J. van Henten, "Data synthesis methods for semantic segmentation in agriculture: a capsicum annuum dataset," 2017, submitted to the Journal of Computers and Electronics in Agriculture.

[6] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *arXiv preprint arXiv:1703.10593*, 2017.

[7] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking Atrous Convolution for Semantic Image Segmentation," *ArXiv e-prints*, June 2017.

[8] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[9] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," in *ICLR*, 2015. [Online]. Available: http://arxiv.org/abs/1412.7062

[10] F. Dittrich, H. Woern, V. Sharma, and S. Yayilgan, "Pixelwise object class segmentation based on synthetic data using an optimized training strategy," in *Networks Soft Computing (ICNSC), 2014 First International Conference on*, Aug 2014, pp. 388–394.

[11] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. Lopez, "The SYNTHIA Dataset: A large collection of synthetic images for semantic segmentation of urban scenes," 2016.

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

[13] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *CoRR*, vol. abs/1611.07004, 2016. [Online]. Available: http://arxiv.org/abs/1611.07004

[14] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385