Semantic UAV Image Segmentation of Mixed Cropping Fields

QUSAI MARASHDEH[1], LUKAS DREES[1] & RIBANA ROSCHER[1]

Abstract: Mixed cropping became an important research topic in the area of sustainable agriculture, aiming on novel insights on how different plants interact with each other in the same agricultural field. Convolutional neural networks show remarkable capabilities to solve tasks like semantic segmentation and have been successfully used for agricultural applications. However, performing such tasks for fields with mixed crops is particularly challenging, especially when only a small number of reference images are available for training and testing neural network models. In this paper, we present a study conducted in the Cluster of Excellence "PhenoRob - Robotics and Phenotyping for Sustainable Crop Production" in which 320 unmanned aerial vehicles (UAVs) images of mixed crop fields from the experimental site Campus Klein-Altendorf are analyzed. Specifically, we perform semantic segmentation using a convolutional neural network with U-Net architecture to distinguish between the mixed crops faba bean and spring wheat. We use two different approaches to create annotation masks which are used to learn our model for segmentation. The goal is to quantify the heterogeneity based on the segmentation results and to analyze the influence of different segmentation masks. We evaluate our results by means of a mean confusion matrix and a visualization of our results comprising the estimated biomass of the two plants and total yield. Preliminary results show an overall accuracy of classified faba beans of 68% and an overall accuracy of classified spring wheat plants of 87%.

1 Introduction

Crop production is an active research area due to the demand of an increased sustainability while ensuring food for a growing population (YADAV et al. 2021). Objectives are focusing on different ways to increase food production considering various aspects such as the limited area, high input cost, saving the soil quality, the effect of pests infestation, and other threats that affect efficient crop production, for instance, climate change. Several studies investigate the statistical analysis of data from intercropping interaction (PEARCE & GRILLIVER 1978). One way is to take advantage of machine learning to analyze observational data to gain insights into the plants' behavior and their interactions. Recently, one successful approach is the use of convolutional neural networks, which have already shown success in related applications such as weed detection or disease detection (KAMILARIS et al. 2018; LOTTES et al. 2018).

In this paper, we present a study conducted within the Cluster of Excellence "PhenoRob - Robotics and Phenotyping for Sustainable Crop Production". In this study, 320 orthophotos were calculated from UAV images acquired over a four-month period from April to July at the experimental field Campus Klein-Altendorf in 2020. The goal of this study is to monitor the mixed plants and derive relevant phenotypic traits such as the yield efficiently at different time points, as shown in Fig. 1. We achieve our goal by performing semantic segmentation using a convolutional neural network

[1] Rheinische Friedrich-Wilhelms-Universität Bonn, Institut für Geodäsie und Geoinformation, Nussallee 15, D-53115 Bonn, E-Mail: qusai.marashdeh94@gmail.com, [ldrees, ribana.roscher]@uni-bonn.de

 DOI: 10.24407/KXP:1796027324

with U-Net architecture to distinguish between the mixed crops faba bean and spring wheat. We use the segmentation results to estimate the heterogeneity from which insights about different mixing ratios can be derived. We face challenges such as a limited spatial resolution, differences in the texture and spectral signature of mixed plants, and the overlap between the mixed components occurring at later growth stages.

Fig. 1: UAV image patches showing the same area of a mixed cropping field with spring wheat and faba bean at different time points: early growth stage (left), intermediate growth stage with which we train our model (middle), and late growth stage (right).

In our experiments, we show that neural networks are able to distinguish between different plants. We evaluate our results in different ways. First, by estimating the confusion matrix for comparing two annotation methods of mixed crop images. Second, we visualize our obtained results with a specific focus on the heterogeneity with respect to the yield.

2 Data

Data was acquired within the currently running cluster of excellence PhenoRob, specifically in the core project ‘New Field Arrangements’, which aims to estimate and evaluate how different crops evolve in mixed cropping fields.

2.1 Study location and sensor measurements

Our study site is located at Campus Klein-Altendorf close to Rheinbach in North Rhine-Westphalia, Germany. The experimental field consists of 320 plots, each of 15 m^2, arranged in 10 rows and 32 columns. On these plots, 8 different wheat varieties are combined with 2 different arable varieties at varying seed density, resulting in 197 mixture plots, 92 spring wheat and 31 faba bean monoculture reference plots. RGB images were acquired with a UAV at approximately 10 meter flight altitude on a weekly basis during the growth period from April to July 2020. From the images, RGB orthomosaics were computed with 3 mm GSD.

2.2 Patch extraction and data annotation

Each of the 320 plots, with the size of 2700 × 480 px is split into 6 quadratic patches of size 450 × 450 px. We annotated 9 mixed crop plots for training and testing our proposed method, where 5 patches from each mixed crop plot are used for training (in total: 45) and one for testing (in total:

9). For all monoculture plots, we performed the annotation semi-automatically in two different ways:
1- Application of vegetation indices (VI): We estimate the RGB vegetation indices (RGBVI) value for each pixel in the image and set an appropriate threshold to differentiate between the vegetation pixels and the background pixels.
2- Application of a pre-trained Gaussian classifier: We use a pre-trained model, which was trained using a set of plant images. The images were collected using two cameras, a Canon 7D and a high-end mobile phone (Sony XPeria Z3 Compact). Twenty images were collected by each of the cameras (40 images overall), capturing images of a fallow field farm in Gatton, Queensland, and garden beds in Brisbane, Queensland as described in (BAWDEN et al., 2017). The images are divided into three sets, 14 images for training, 14 images for validation, 12 for testing. We apply it to our data to differentiate between vegetation and background.
For images with mixed crops, we annotate in 2 steps: First, we annotate faba beans manually and second, we use the described procedure for monoculture images to annotate spring wheat plants. We combine both annotations by overlaying VI respectively Gaussian derived masks with the manual labeled masks as illustrated in Fig. 2.

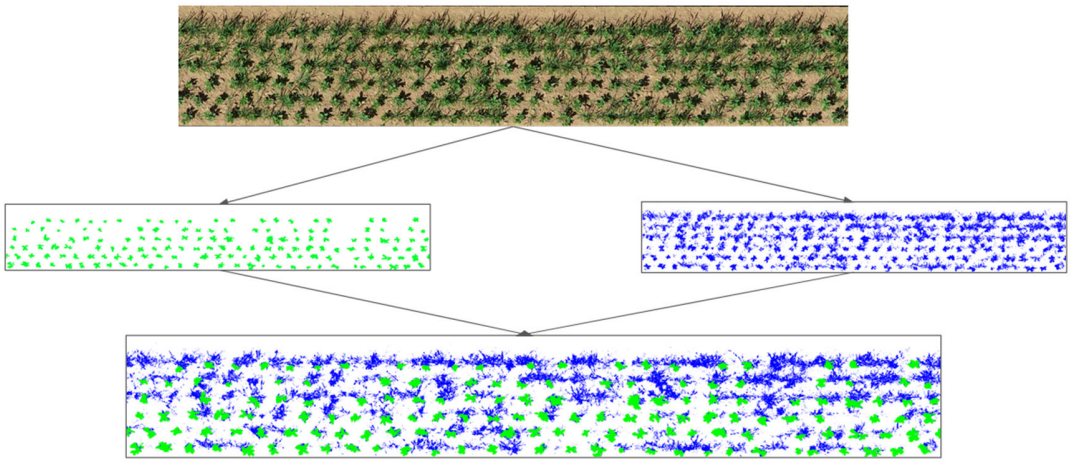

Fig. 2: Steps of producing the label mask. Top: Original image. Middle right: Resulting mask using a pre-trained Gaussian classifier for spring wheat. Middle left: Manually annotated mask for faba bean. Bottom: Combined mask for faba bean and spring wheat.

3 Methodology

We use U-Net architecture (RONNEBERGER et al. 2015) which already showed success for similar applications, and the number of parameters is comparably low.

3.1 U-Net: Convolutional network architecture

The architecture of the U-Net is illustrated in Fig. 3. Our implementation was performed with PyTorch in Python, which is an open-source library based on the Torch library.
The architecture contains two paths: encoding path and decoding path. In the encoder, a lower-dimensional representation of the data is computed to capture the most relevant patterns, while in

the decoder, the representation is mapped back to an intended outcome such as a segmentation mask. The intermediate results, which are obtained after each operation in the network are called feature maps.

1- Encoding path: This part consists of four repeated blocks, each containing two 3×3 convolutions with a rectified linear unit (ReLU) activation applied after each convolutional operation and a 2×2 max pooling operation for downsampling purposes. In each block, we double the number of convolutional filters in comparison to the previous block.

2- Decoding path: The decoder has a symmetric structure to the encoder, however, instead of downsampling operations we use upsampling operations. The final outcome has the same number of rows and columns as the input image. Each block contains an upsampling of the feature map with a 2×2 convolution, where the number of convolutional filters is reduced by 2 in comparison to the previous block. The architecture contains skip-connections, which means that the output of a previous block in the decoder is concatenated with the corresponding feature map in the encoder before applying the operations in each block, which are two 3×3 convolutions with ReLU after each of them. In the final layer, we use a 1×1 convolution layer to project the input channels to the desired output channels according to the class number, and a softmax activation function.

3.2 Evaluation metrics

We evaluate the segmentation results in two different ways. First, by means of a confusion matrix containing the three classes spring wheat, faba bean, and background. The matrix estimates the deviation between the reference mask and the predicted one. Moreover, we present the average confusion matrix. Further, we show the correlation of our estimation of the heterogeneity of the mixed plants by counting the number of pixels of each class and the total yield for each class.

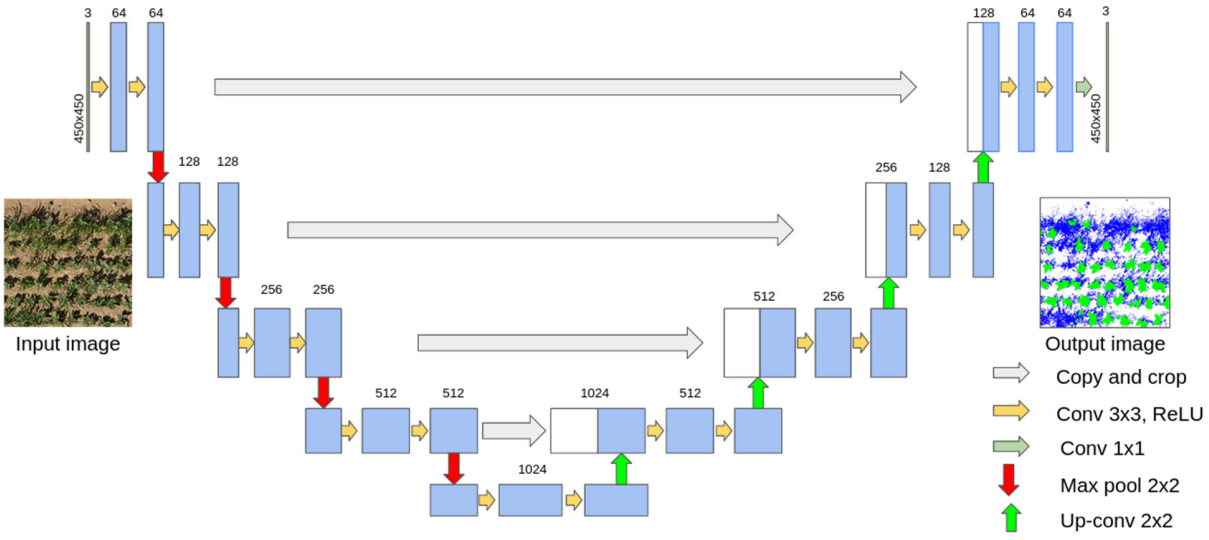

Fig. 3: U-Net architecture with skip connections (gray arrows) used for semantic segmentation. The architecture consists of an encoder (left part of the architecture) and a decoder (right part of the architecture). All arrows represent different mathematical operations.

4 Experiments

4.1 Experimental setup

We perform various experiments to evaluate (1) the influence of different annotation methods, (2) the capability to apply our model to monoculture images of the mixture components, and (3) how close we get with the mixture ratio derived from the segmentation images to in-situ reference measurements performed in the field.

For training the U-Net, we use cross-entropy loss with an Adaptive Moment Estimation (Adam) optimizer and a batch size of 1. Using one-side padding allows us to get an output image size equal to the input size. As data augmentation methods, random horizontal and vertical flipping as well as cropping are used, which help to avoid overfitting. Using all 45 mixed image patches of the training set, the training starts to converge after 200 epochs and thus takes about 2.5 hours. Since there are many monoculture images available that are easy to annotate, in initial experiments, we used an extended training dataset that includes both annotated monocultures and mixed cropping images. However, it has been found that the convergence time takes longer without improving the model, so the monocultures are only used for evaluation.

4.2 Results and Discussions

In Fig. 4, we first compare the confusion matrices of the two annotation methods using VI and a pre-trained Gaussian classifier, where the matrices are split into the classes background, faba bean (FB), and spring wheat (SW). The highest score in both cases is for the class background, which is reliably segmented in more than 95%. This is followed by SW, which is captured slightly better by the model trained using VI (92% compared to 87%). In return, the Gaussian trained model is 8 percentage points better for FB, although at a lower level than SW (68% to 58%). For the Gaussian results, we notice that most of the incorrectly predicted SW pixels are labeled as FB (11%), and vice versa most of the incorrectly predicted FB pixels are labeled as SW (20%). For VI results, the incorrectly predicted pixels of the SW class are equally often confused with FB and background. Noticeably, the incorrectly predicted FB pixels are not equally distributed among the other classes, but mainly belong to SW (34%).

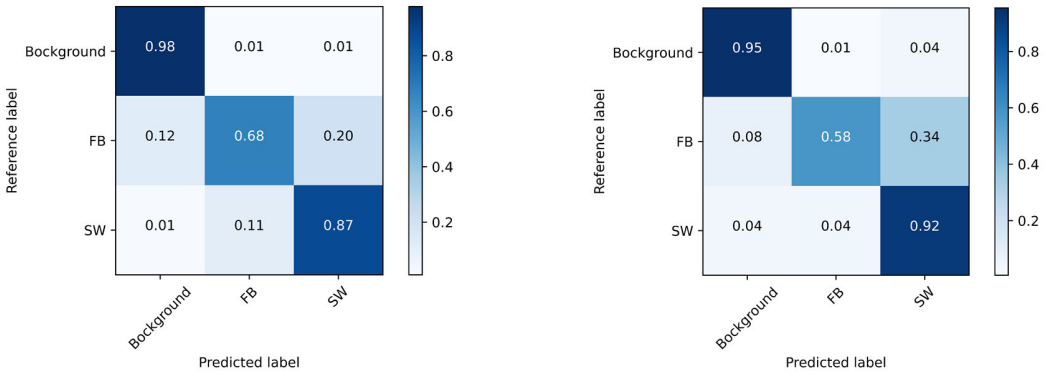

Fig. 4: Confusion matrix of semantic segmentation using the result of Gaussian annotation results (left) and VI annotation results (right) of training only 45 mixed patches to train our U-Net model (%).

Fig. 5 visualizes the differences of the annotation method on an exemplary mixture test patch. It is noticeable that the amount of annotated SW pixels varies significantly. While the result with the pre-trained Gaussian classifier contains less SW overall, the VI result contains more. Accordingly, this trend can also be seen in the segmentation. Although VI masks are controllable via selected threshold value, it is difficult to avoid annotating shadow areas: In the case of a too low threshold value, shadow areas are avoided, but a large number of holes on the plants are created. As a consequence, we have chosen a rather large threshold value in order to capture the plants completely, resulting in some false vegetation pixels in soil and shady areas than with the Gaussian classifier.

We assume that the high threshold value contributes to SW having a higher accuracy with VI annotation. In contrast, one can see in the images that FB annotations are less suppressed in the Gaussian result, resulting in thicker segmentation for this approach. Remarkably, with both annotation methods, our model succeeds in detecting FB plants that were missed in the manual annotation or are heavily overgrown with wheat (top field-row in the images).

To assess the stability of our model, we also applied the model to the monoculture images without training on them before. Fig. 6 shows an example of the results of the VI trained model on both types of monocultures, SW and FB. The results look promising: There are only minor outliers in the FB monoculture, while there are no FB outliers in the SW. It is worth noting that the model successfully segments contiguous FB plants, although the annotation is based on free-standing FB plants in the mixtures only.

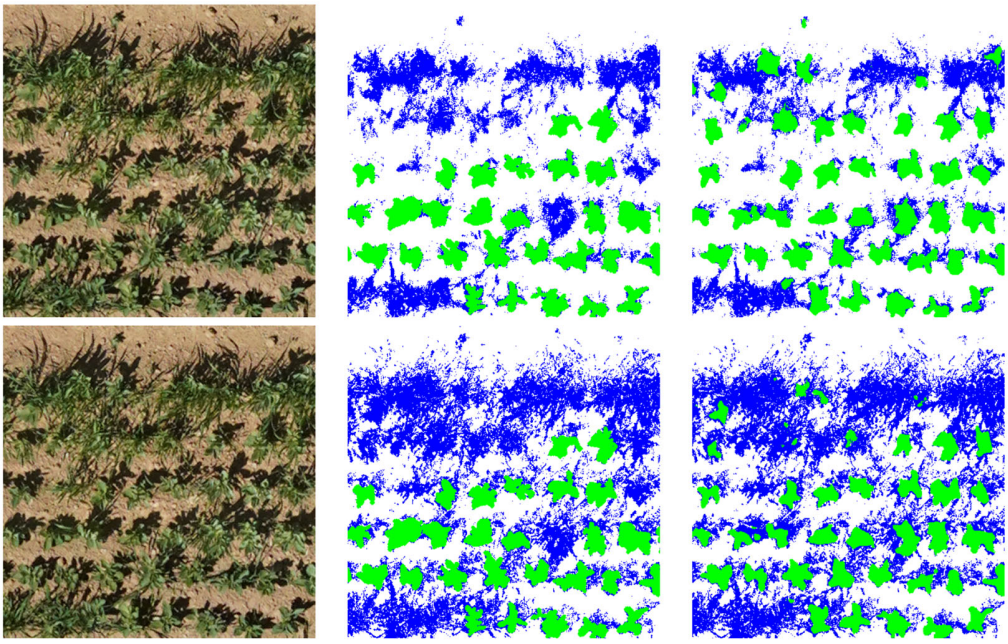

Fig. 5: Segmentation results of a mixture test patch. The first row shows the segmentation result using the annotation obtained from the pre-trained Gaussian classifier. The second row shows the segmentation result using VI for training. From left to right: Image patches, reference annotation masks, and predicted segmentation masks.

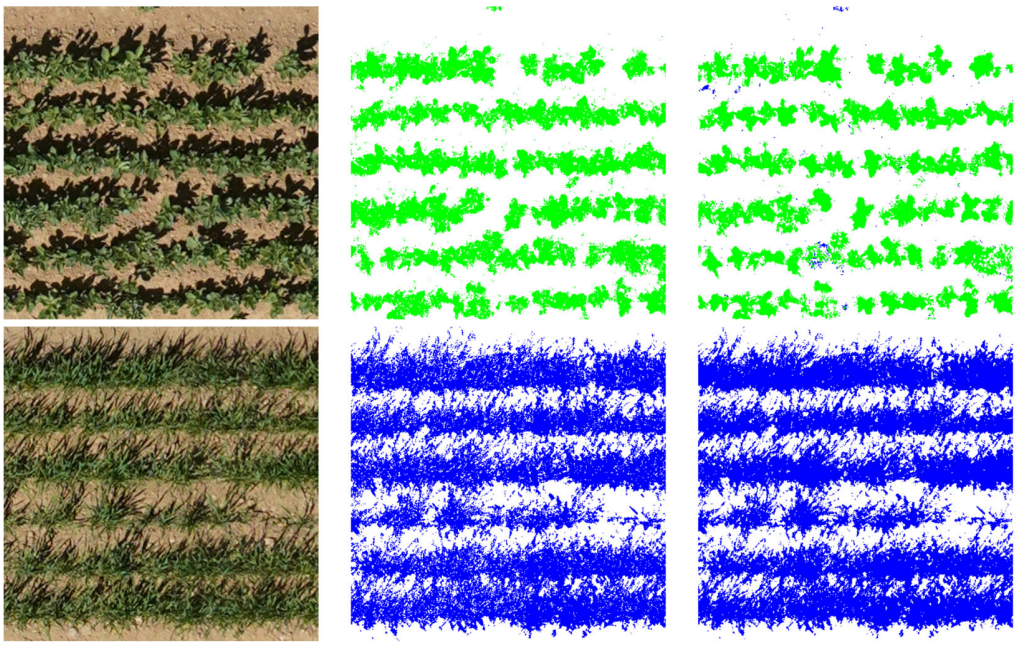

Fig. 6: Segmentation results of faba bean (top) and spring wheat (bottom) monocultures using VI-based U-Net training. From left to right: Image patches, reference annotation masks, and predicted segmentation masks.

To highlight the practical utility of the segmentations, we continue to use the VI model and intend to compare the results with in-situ measurements from the field. Approximately four weeks after the date the images were taken, the yield of the mixture components was manually measured in the field. With this, we assume a correlation between yield in the mixture and the mixing ratio, which we derive from the pixel ratio of the classes SW and FB in the segmentation image.
Fig. 7 shows the R^2 scatter plots in which the in-situ yield in gram (y-axis) of all mixture test images is compared to the respective FB and SW ratios derived from the images (x-axis). Thus each dot represents an image, and the R^2 value (maximum 1) indicates how well the regression line is determined by the dots. For both FB and SW, the R^2 values are rather low, but there is a clear trend indicating that a higher ratio also results in a correspondingly higher yield. Besides, the R^2 plot can only be an approximation, on which two factors have a major influence. First, the temporal difference between the in-situ data and our model estimates, because the mixing ratio may still have developed differently after the image-taking date. Second, the perspective, as the 2D image only considers the canopy surface, whereas 3D dimensional measurements can also count yields of beans and wheat grains that are located between the soil and the upper canopy. Nevertheless, the result indicates that our model is well suited to perform semantic segmentation in complex mixed cropping systems and can provide reasonable approximations for important phenotypic traits, such as the mixture ratio and yield.

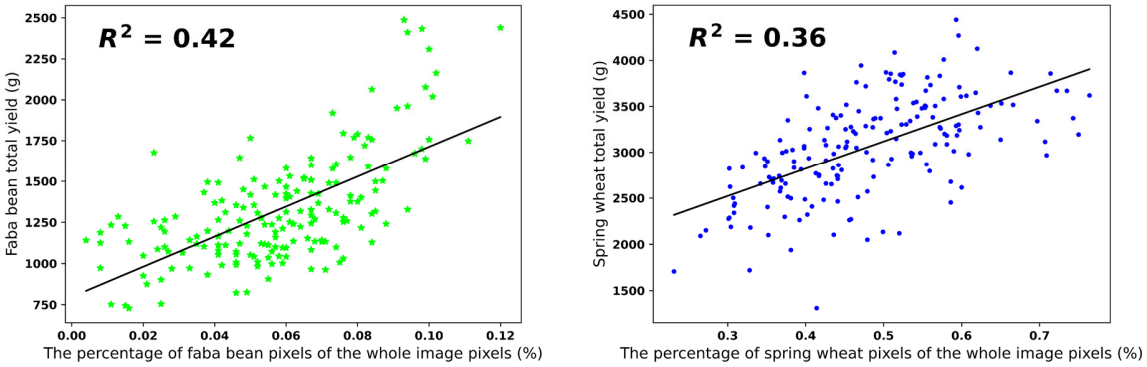

Fig. 7: The correlation between in-situ measured yield in gram (y-axis) and from segmentation images derived mixing ratio in percentage (x-axis) for faba bean on the left and spring wheat on the right. Segmentation images were obtained using VI-based U-Net training.

5 Conclusion

In this paper, we demonstrate a novel application of semantic RGB image segmentation using deep learning to highly complex mixed cropping environments. For this purpose, we use a neural network with the U-Net architecture, which is widely used in the field of semantic segmentation. Since it is time-consuming and difficult to perform manual annotations on such finely-structured and overlapping spring wheat and faba bean plants of a mixed cropping image, we compare two approaches for annotation generation, namely vegetation index-based annotation and annotation by Gaussian classifier. This allows us to manually annotate only the beans from the composite of mixed cropping partners, while wheat is labeled automatically.

Our experiments show that at a medium growth stage we are able to segment wheat in mixtures with up to 95% and beans with up to 70%. It turns out that the choice of both annotation methods has benefits and drawbacks. The VI-based model over-interprets the vegetation because a rather high threshold is needed to close holes in the canopy, but thus shaded areas are also annotated. Hence, spring wheat is well segmented, while beans are often incorrectly segmented as wheat. Meanwhile, the Gaussian classifier provides a stronger separation between vegetation and background, resulting in a better segmentation result for beans, but less correctly segmented wheat. We further show that our model is able to segment monoculture images of mixing components with satisfactory results. Comparison of the mixing ratio derived from the image with in-situ measurements taken later in the field shows trends that automatized image segmentation using deep neural networks is useful in areas of mixed cropping to approximate agricultural yield.

6 Acknowledgment

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2070 – 390732324.

7 References

BAWDEN, O., KULK, J., RUSSELL, R., MCCOOL, C., ENGLISH, A., DAYOUB, F., LEHNERT, F. & PEREZ, T., 2017: Robot for weed species plant-specific management. Journal of Field Robotics, **34**(6), 1179-1199, https://doi.org/10.1002/rob.21727.

KAMILARIS, A. & PRENAFETA-BOLDÚ, F.X., 2018: Deep learning in agriculture: A survey. Computers and electronics in agriculture, **147**, 70-90.

LOTTES, P., BEHLEY, J., MILIOTO, A. & STACHNISS, C., 2018: Fully convolutional networks with sequential information for robust crop and weed detection in precision farming. IEEE Robotics and Automation Letters, **3**(4), 2870-2877, https://10.1109/LRA.2018.2846289.

PEARCE, S.C. & GILLIVER, B., 1978: The statistical analysis of data from intercropping experiments. The Journal of Agricultural Science, **91**(3), 625-632.

RONNEBERGER, O., FISCHER, P. & BROX, T., 2015: U-Net: Convolutional Networks for Biomedical Image Segmentation. CoRR, abs/1505.04597, http://arxiv.org/abs/1505.04597.

YADAV, P., JAISWAL, D.K. & SINHA, R.K., 2021: Climate change: Impact on agricultural production and sustainable mitigation. S. Singh, P. Singh, S. Rangabhashiyam, K. Srivastava (eds), Global Climate Change, Elsevier, 151-174.