Joint Plant Instance Detection and Leaf Count Estimation for In-Field Plant Phenotyping

Jan Weyler, Andres Milioto, Tillmann Falck, Jens Behley, and Cyrill Stachniss

***Abstract*—Precision management of agricultural fields as well as plant breeding are central factors for keeping yields high and to provide food, feed, and fiber for our society. A key element in breeding trials but also for targeted management actions is to analyze the growth state of individual plants objectively and at a large scale. In this paper, we address the problem of analyzing crops in real agricultural fields based on camera data recorded with mobile robots and to derive information about the plant development, e.g., to monitor phenotypic traits such as growth stage. We propose a novel single-stage object detection approach that localizes crops and weeds in the field. At the same time, it detects plant-specific leaf keypoints intending to estimate leaf count at a plant level, which is a key trait for classifying the growth stage. We implemented and thoroughly tested our approach on real sugar beet fields. As our experiments show, it performs the required detections and shows superior performance with respect to a state-of-the-art two-stage approach based on Mask R-CNN.**

***Index Terms*—Robotics and Automation in Agriculture and Forestry, Deep Learning for Visual Perception**

I. Introduction

CROP production provides food, feed, and fiber for our society. To keep the yields high and to adapt to stresses as well as to impacts of climate change, plant breeders continuously generate new genetic variations of crops. These variations are then planted and their performance is assessed. Thus, plant breeders are looking for effective systems to assess detailed phenotypic traits about plants at a large scale for an in-depth understanding of the relationship between genotype and phenotype [5]. Regular and standardized monitoring of the plant's vegetative development is required by law in many countries to maintain high-quality seeds of good varieties [4]. Recording how the individual plants develop and grow however is a time-consuming process and is conventionally done, at least in the field, by manual inspection. Crops that show desirable traits then serve as the basis for the next round when generating genetic variations.

Recognizing the growth state of individual plants is also relevant for targeted management actions on agricultural fields performed by autonomous robots to ensure effective weed control and crop safety. For example, most weeds show the highest susceptibility to herbicides at certain growth stages [17]. However, when applying postemergence herbicides it is also important to consider the crop growth stage to avoid potential crop injury [16]. Thus, agricultural robots must identify the current state of the plant development to make informed decisions.

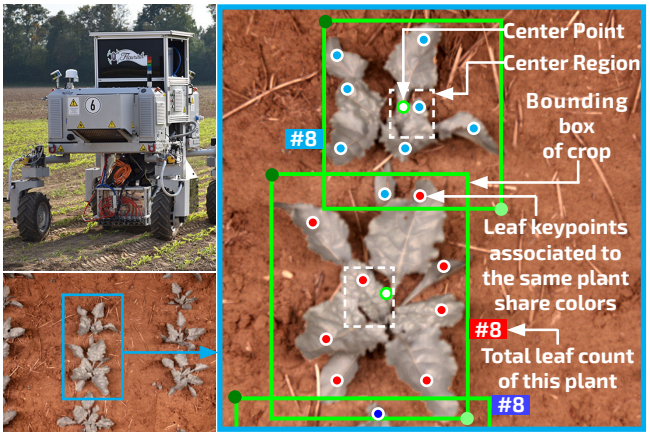

Fig. 1. Left: Agricultural robot (top) acquires image data (bottom) which is fed as input to our system. Right: Magnified view of the result of our vision-based system. We predict the bounding boxes of crops (green) and weeds by their corresponding top-left (dark green circle) and bottom-right (light green circle) corner. Simultaneously we associate plant-specific leaf keypoints to each plant to estimate its total leaf count (#) in complex scenes.

Vegetative development stages such as the BBCH index [14] are mainly defined by the number of leaves produced on the main stem and the number of tillers on a plant (e.g., cereals), or the number of nodes on a plant (broadleaf plants). Thus, the leaf count is a key plant trait and is directly related to the growth stage of the plant [3], its yield potential [32] and proper herbicide timing [33]. Today, the vegetative stage is obtained by manual inspection, typically sampled at a subset of locations in the field. By contrast, automating this process allows for a more frequent assessment on a large scale in less time to support plant breeders and precision farming on agricultural fields.

In this paper, we address the problem of robotic in-field analysis based on images to derive information about the vegetative stage of each plant using leaf information. Combining such an approach with unmanned aerial vehicles or ground robots allows for analyzing breeding plots and agricultural fields on a large scale at a high time resolution. Using data from the field instead of the laboratory is important for phenotyping since the phenotype is a result of genetic

Manuscript received: October, 15, 2020; Revised: January, 8, 2021; Accepted: February, 2, 2021. This paper was recommended for publication by Editor Youngjin Choi upon evaluation of the Associate Editor and Reviewers' comments.

This work has partially been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy, EXC-2070 - 390732324 - PhenoRob and the Robert Bosch GmbH.

J. Weyler, A. Milioto, J. Behley, and C. Stachniss are with the University of Bonn, Germany. T. Falck is with Robert Bosch GmbH, Germany. jan.weyler@igg.uni-bonn.de

Digital Object Identifier (DOI): see top of this page.

expression, environmental influences, and field management.

The main contribution of this work is a single-stage object detection approach based on CenterNet [7] that can jointly localize crops and weeds and detect plant-specific leaf keypoints using images. Our approach is applicable to real-world field data, see Fig. 1, and provides leaf counts for every plant in the field. It shows superior performance with respect to an alternative state-of-the-art two-stage approach based on Mask R-CNN [12].

In sum, we make the following four key claims. First, our method is able to accurately detect crops and weeds without relying on expensive anchor-based frameworks. Second, it simultaneously detects leaf keypoints associated with individual plants enabling per plant leaf counts in complex scenes with overlapping plants. Third, in contrast to related methods, we do not rely on any a priori assumption about the number of keypoints associated with each object. Fourth, this allows for the computation of relevant basic phenotypic traits on real-world field data in an automated fashion, e.g., to support plant breeders and precision farming on agricultural fields.

II. Related Work

The development of vision-based systems for plant classification [27], detection [24] and extraction of morphological plant traits [9] has made significant progress in recent years. Some approaches rely on a set of handcrafted features [34], but more recent methods use convolutional neural networks (CNNs), which learn features directly from labeled image data [28].

Crop-Weed Classification/Segmentation. In the context of crop-weed classification, several vision-based approaches have been proposed [2], [23]. Lottes *et al*. [25] propose a two-stage approach to distinguish crops and weeds in the field based on RGB and near infra-red (NIR) images. They identify vegetation using NIR images followed by a random forest classifier based on a set of handcrafted features.

In contrast, more recent methods use CNNs to distinguish crops and weeds. Milioto *et al*. [27] perform real-time classification of sugar beets and weeds by detecting individual plants as connected components and feed each of them to a CNN to predict the plant type. McCool *et al*. [26] employ model compression on a complex CNN to conduct an ensemble of lightweight CNNs to address the task of crop and weed segmentation. Lottes *et al*. [23] perform end-to-end semantic segmentation to detect crops and weeds as well as stem positions based on RGB and NIR images to enable farming robots to apply selective weed treatment. Unlike our approach, they propose a pixel-wise semantic classification of the entire image but do not detect individual plants as an instance itself, which is necessary to predict phenotypic traits of single plants. Bargoti *et al*. [2] develop an image-based fruit detection system to support yield mapping and robotic harvesting. Their approach is based on Faster R-CNN [29], a two-stage object detection framework, to localize each fruit by its bounding box.

Leaf Counting. In contrast to the above-mentioned approaches, current leaf counting methods are mainly conducted in laboratory environments [30] based on cropped images or 3D models of single plants but are often not applied in real agricultural fields. Aksoy *et al*. [1] propose a clustering algorithm to extract leaves in NIR images of single plants. Since the output given by the clustering approach splits one leaf into more than one segment and may contain noisy regions they merge initial segments based on a leaf-shape descriptor represented by the convex hulls of segments. Kumar *et al*. [18] present a graph-based algorithm to segment leaf regions in enhanced HSV images of single plants. Given the assumption of round leaves, they predict the total count by applying a circle Hough transform. In contrast, Golbach *et al*. [11] propose a segmentation method based on a flood-fill algorithm to identify leaves from 3D models of single plants, which enables them to measure the area of singles leaves.

More recent methods use deep neural networks for leaf count estimation. The approach by Itzhaky *et al*. [15] computes the number of leaves from multiple image scales of single plants accounting for cases of small and large leaves using a feature pyramid network [20]. They predict a heatmap of leaf keypoints and feed this map to a non-linear regression model to estimate the total leaf count. Shi *et al*. [31] propose a multi-view 3D segmentation approach to segment point clouds of single tomato seedlings acquired in a laboratory setting into leaves, stems, and nodes. They project predictions of a neural network applied to multiple 2D images into a 3D point cloud to eliminate errors and improve performance.

In contrast, our proposed pipeline bridges the gap of separate crop-weed detection and leaf counting by performing both tasks simultaneously in an end-to-end manner on images of agricultural fields. This task is more challenging in comparison to laboratory settings since different plants may overlap. Our method enables high-throughput phenotyping on real-world field data and can be integrated into an autonomous robotic platform.

Our approach is related to the task of human pose estimation proposed by Zhou *et al*. [35]. They predict the bounding box of each person in an image and simultaneously estimate the location of each joint via heatmaps and offsets to its corresponding center point. However, their method relies on a fixed number of keypoints for each object, which is assumed to be known a priori. This assumption is not valid for plants since the number of leaves per plant is highly variable. In contrast, our approach is able to associate a varying number of leaf keypoints to each plant to account for different plant growth stages.

III. Our Approach

The main objective of our work is to enable agricultural robots to detect and simultaneously distinguish crops and weeds and estimate their leaf count in order to assess the vegetative development of single plants in an automated fashion. We propose a vision-based approach for the joint processing of crop-weed detection and leaf count estimation based on a fully convolutional neural network (FCN) [22].

Given an image of the field, we feed it to a feature pyramid network (FPN) [20] using a ResNet18 model [13] pre-trained

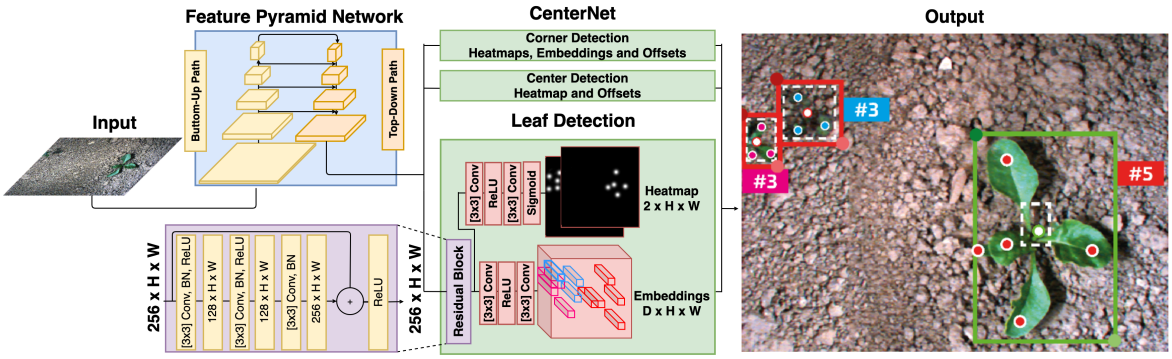

Fig. 2. Network architecture used for joint crop weed detection and leaf count estimation. We predict the bounding boxes of crops (green) and weeds (red) based on CenterNet and simultaneously detect the location of leaf keypoints via heatmaps, which depth-dimension encodes crops and weeds. Additionally, we predict embeddings for all leaf keypoints to associate each to an individual plant on the field to estimate its total leaf count. (Best viewed in color.)

on ImageNet [6] to construct a high resolution semantic feature volume. We use the output of the FPN to first detect the bounding boxes of crops and weeds using an FCN based on the recently proposed CenterNet [7] approach (Sec. III-A). We designed a new version of CenterNet, which allows us to jointly detect plant-species-specific keypoints inside each leaf (Sec. III-B) and associate them to each detected instance (Sec. III-C), enabling us to finally determine the total leaf count of individual crops and weeds, see Fig. 2.

A. Crop-Weed Detection

The backbone of our proposed network is CenterNet [7], which detects each object as a triplet of keypoints, i.e., its top-left corner, bottom-right corner, and center point.

This object detection pipeline predicts the location of all object's top-left and bottom-right corners in an image and determines corresponding pairs that belong to the same object to compute bounding boxes. To reduce the number of false positive detections CenterNet additionally detects the center points of each object's bounding box and preserves only those previously determined boxes which contain a center point inside a defined central region.

To predict the corner locations, the network computes two heatmaps of top-left and the bottom-right corners for all crops and weeds separately. Those heatmaps contain high confidence scores at corner locations, see Fig. 3. Simultaneously, the network estimates the location of all center points in an image by an additional heatmap for both categories.

As a result of down-sampled heatmaps, a remapping of the detected corner and center coordinates to the input image leads to coarse predictions. To alleviate this problem, the network additionally predicts offsets for all corners and centers for more precise localization of these points.

To match corresponding corners from the same object, the network also computes an embedding vector for each corner in the heatmaps. We train the network with the objective of preserving small distances between embeddings of the same object, and large distances for different instances. Thus, we associate top-left and bottom-right corners of the same object by computing their distance in an automated post-processing step on top of the network predictions.

Post-processing. To generate bounding boxes from the heatmaps, embeddings, and offsets, an automated post-processing procedure after the CNN is applied. First, we adopt non-maximum suppression (NMS) by using a max pooling layer with kernel size k_{NMS} on the heatmaps and keep only those keypoints whose value is identical to its original value to remove redundant corners and centers. Second, we select the top-k predictions of each heatmap according to the confidence score. Third, we shift the selected top-left (tl) and bottom-right (br) corners as well as center points by their corresponding offsets. Finally, we pair selected corners if the $L1$-distance of their associated embeddings is less than a predefined threshold θ_{co} to determine bounding boxes. During pairing, we only consider corners of the same category where $x_{\text{br}} > x_{\text{tl}}$ and $y_{\text{br}} > y_{\text{tl}}$. Note that each top-left corner has a maximum of j_{max} related bottom-right corners and vice versa selected by shortest distance.

Quantitative results, however, suggest a high rate of false positive predictions, since visual patterns inside bounding boxes are not exploited by corners [7]. This problem is solved by including the adjusted center point predictions. First, each previously generated bounding box is associated with a scale-aware center region, see Fig. 3. A bounding box is valid if and only if a center point of the same category is detected within this region, otherwise it is discarded. We keep only those boxes whose associated triplet of keypoints have each a confidence score higher than a predefined threshold θ_{score}. Finally, we compute the confidence score of an object by averaging the predicted confidence scores of its triplet.

B. Leaf Keypoint Detection

Simultaneously to the aforementioned object detection pipeline, we predict morphological plant traits of each detected plant, i.e., its total number of leaves and their location. We propose a counting by detection approach, detecting leaf keypoints encoded in a 2-channel heatmap of size H×W, where H and W specify the height and width accordingly. Each channel predicts confidence scores of crop and weed leaf keypoints separately, due to different visual appearance. Thus, we encode the location of plant-species-specific leaf keypoints in each channel, shown as white circles in Fig. 2.

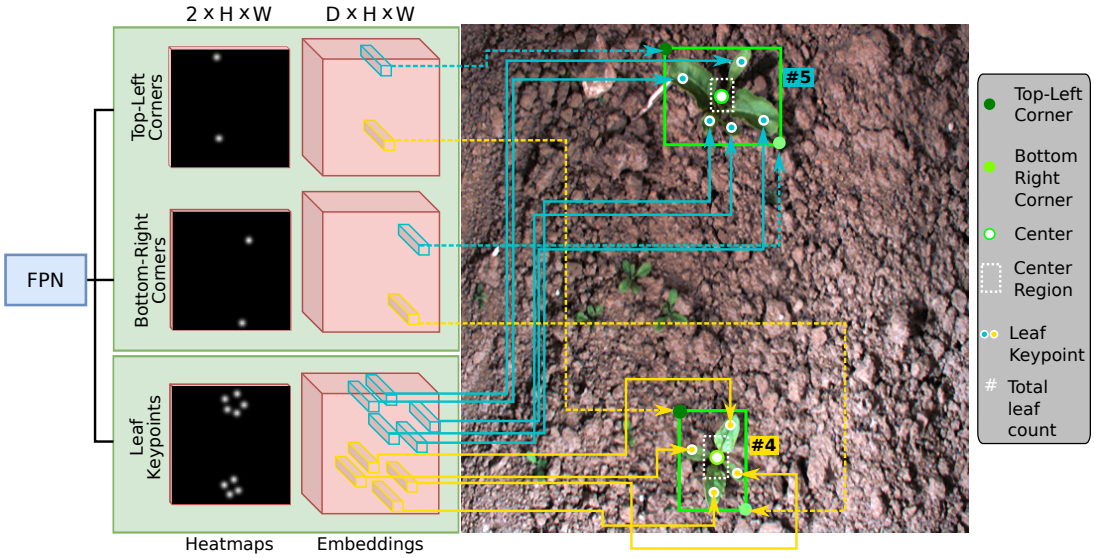

Fig. 3. The network predicts heatmaps, encoding the location of corners and centers (not shown) of bounding boxes. An additional heatmap estimates the location of leaf keypoints. For each detected corner and leaf keypoint we predict an embedding vector to associate pairs of the same plant. Thus, we match top-left and bottom-right corners based on their distance to compute bounding boxes and assign each leaf keypoint to a specific plant based on its embedding as well. We visualize associative embeddings of the same plant in uniform colors. Here we show only the predictions of crops.

During training, we create a map of weights w of size $2\times\mathrm{H}\times\mathrm{W}$ with unnormalized, isotropic 2D Gaussians, i.e., $\exp\left(-\frac{1}{2}(\Delta x^2+\Delta y^2)\,\sigma_{\mathrm{LK}}^{-2}\right)$, placed at each ground truth leaf keypoint to reduce the penalty given to negative locations close by the desired location. Let Δx and Δy be the offsets between the desired location to each pixel in the map w. We choose a fixed radius r_{LK} for each 2D Gaussian such that nearby leaf keypoints do not intersect within the interval defined by r_{LK} and set $\sigma_{\mathrm{LK}}=\frac{1}{3}r_{\mathrm{LK}}$. Thus, we argue that predictions close to their desired ground truth location might still detect a valid leaf keypoint. Given the predicted heatmap, we apply a per-pixel sigmoid and define $\mathcal{L}_{\mathrm{det}}^{\mathrm{leaves}}$ as a variant of focal loss [19], [21] to encounter class imbalance:

$$\mathcal{L}_{\mathrm{det}}^{\mathrm{leaves}}=-\frac{1}{N_L}\sum_{c=1}^{2}\sum_{i=1}^{H}\sum_{j=1}^{W}w_t^{\beta}\left(1-p_t\right)^{\gamma}\log\left(p_t\right), \quad (1)$$

with

$$p_t=\begin{cases}p_{cij} & \text{if } y_{cij}=1\\ 1-p_{cij} & \text{otherwise}\end{cases}, w_t=\begin{cases}w_{cij} & \text{if } y_{cij}=1\\ 1-w_{cij} & \text{otherwise}\end{cases},$$

where N_L is the total number of leaves in an image. Let $p_{cij}\in[0,1]$ be the model's estimated probability for a leaf keypoint at location (i,j) in the c-th channel encoding crops or weeds respectively and $y_{cij}\in\{0,1\}$ is the ground-truth category. We set the hyperparameters γ and β to 2 and 4 following Duan *et al.* [7], where β controls the penalty given to negative locations nearby the ground truth leaf keypoint.

We initialize the biases of all convolutional layers to zero expect for the final layer, which we initialize to $b=-\log((1-0.01)/0.01)$ due to class imbalance [21]. Thus, the probability of positive samples is small compared to those of negative samples in the first iterations of training to optimize a smaller and more stable loss.

C. Associating Leaf Keypoints

In laboratory settings, plants are observed instance-wise and therefore detection of leaf keypoints is in general sufficient to count the number of leaves [10], [15]. However, these approaches fail in real-world fields, since each image contains multiple plants that may even overlap, see Fig. 4. To associate each detected leaf keypoint to a specific plant on the field we additionally predict embedding vectors for each leaf keypoint and train the network with the objective of preserving small distances to their associative pair of top-left and bottom-right corner. Thus, we can assign each detected leaf keypoint to an individual plant by distance based on an automated post-processing procedure, see Fig. 3.

We achieve this objective by the following proposed clustering loss function. We separate our loss function into two parts. First, we pull spatial embeddings of leaf keypoints belonging to the same instance:

$$\mathcal{L}_{\mathrm{pull}}^{\mathrm{leaves}}=\frac{1}{P}\sum_{k=1}^{P}\frac{1}{N_P}\sum_{i=1}^{N_P}\left(\bar{e}_{tb,k}-e_{L_i,k}\right)^2, \quad (2)$$

where P is the total number of plants in the image and N_P is the number of leaves associated with each plant. Let $\bar{e}_{tb,k}$ be the average embedding of the top-left and bottom-right corner of the k-th plant and $e_{L_i,k}$ be the embedding of the i-th leaf keypoint associated with this plant.

Second, we push apart spatial embeddings of leaf keypoints belonging to different instances and leaf keypoints to non-associative pairs of top-left and bottom-right corner:

$$\mathcal{L}_{\mathrm{push}}^{\mathrm{leaves}}=\frac{1}{P(P-1)}\sum_{k=1}^{P}\sum_{\substack{j=1\\ j\neq k}}^{P}\delta^{\mathrm{leaves}}+\delta^{\mathrm{tb}}, \quad (3)$$

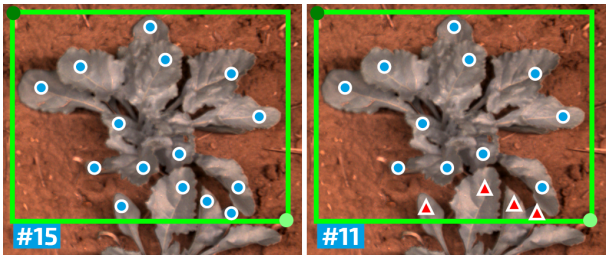

Fig. 4. Example of different leaf counting approaches. Left: Estimating the total number of leaves by counting the predicted leaf keypoints (blue circles) within an object's bounding box. However, this method fails in the case of overlapping plants in real agricultural fields. Thus, we associate each predicted leaf keypoint to an object's top-left and bottom-right corner via embeddings. This allows us to discard leaf keypoints that belong to different plants (red triangles) to ensure more accurate leaf counts (#).

with

$$\delta^{\text{leaves}} = \max\left(0, \Delta - |\bar{e}_{L,k} - \bar{e}_{L,j}|\right), \tag{4}$$

$$\delta^{\text{tb}} = \max\left(0, \Delta - |\bar{e}_{L,k} - \bar{e}_{tb,j}|\right), \tag{5}$$

where $\bar{e}_{L,k}$ is the average embedding of all leaf keypoints associated with the k-th object in the image. The objective $\mathcal{L}_{\text{push}}^{\text{leaves}}$ is high if the embedding $L1$-distance of different objects is smaller than Δ. By minimizing this loss we push apart embeddings of non-associative objects in feature space.

We do not predict offsets for leaf keypoints, since their exact location is not required for the task of counting.

Post-processing. To assign each detected leaf keypoint to an individual plant, an automated post-processing procedure after the CNN is applied. First, we apply NMS on the predicted heatmap to prevent over-counting keypoints as described in Sec. III-A and select its top-n predictions by score. For each leaf keypoint, we compute the $L1$-distance between its corresponding embedding and the averaged embedding pair of top-left and bottom-right corner for each bounding box. We assign a leaf keypoint to its nearest neighbor if the distance is less than a predefined threshold θ_{leaf}. Furthermore, we discard leaf keypoints associated with a plant of a different category or if it is not within the bounding box. Finally, we compute the total number of leaves per plant by counting its associated leaf keypoints, as illustrated in Fig. 4.

IV. Experimental Evaluation

The experiments show the capabilities of our method and support our claims made in the introduction that our system accurately detects crops and weeds without relying on expensive anchor-based frameworks and simultaneously detects leaf keypoints without any a priori assumptions which we associate to specific plants to determine per plant leaf count in complex scenes with overlapping plants in agricultural fields.

A. Experimental Setup

We evaluate on images of a sugar beet field located near Stuttgart in Germany. The dataset contains 705 images of sugar beets and different weed types in a variety of growth stages with a size of $2048 \times 1536\,\text{px}$ and a ground sampling distance of $0.7\,\frac{\text{mm}}{\text{px}}$. The image data was recorded with an agricultural robot equipped with a Manta G-319 GigE camera mounted in nadir view with an additional triple bandpass filter. Note that the bandpass filters maps green, red and NIR. We determine the development under different plant sizes subdividing into them into small (area $< 5\,\text{cm}^2$), medium ($5\,\text{cm}^2 \leq$ area $< 45\,\text{cm}^2$) and large (area $\geq 45\,\text{cm}^2$), see Tab. I. In the following, they are denoted with S, M, and L, respectively. For the training of our network, we use $70\,\%$ of the entire dataset and $10\,\%$ to validate the hyperparameters. To evaluate the final metrics, we rely only on the remaining test portion.

TABLE I
STATISTICS OF OUR TEST DATASET

		all	**small** (S)	**medium** (M)	**large** (L)
sugar beets	**count**	1782	32	187	1563
	min leaves	1	1	1	1
	max leaves	16	3	8	16
	max IoU	0.32	0.13	0.23	0.32
weeds	**count**	836	239	457	140
	min leaves	1	1	1	1
	max leaves	44	5	10	44
	max IoU	0.17	0.14	0.17	0.14

We evaluate the performance of our crop-weed detection in terms of average precision (AP) and average recall (AR) [8], calculated across both categories and on a per-category basis by setting the threshold for intersection over union (IoU) with ground truth bounding boxes at different levels for each of the above-mentioned object scales, i.e., $IoU \in [0.5, 0.95]$ with step size 0.05. To evaluate leaf count performance, we follow the evaluation metrics proposed by Scharr *et al.* [30]:

- *Difference in Count* (*DiC*): mean and standard deviation of the difference between actual and predicted count.
- *Absolute Difference in Count* (*aDiC*): identical to *DiC* but instead computes the absolute value.
- *Percentage Agreement* (*Pa*): number of times the predicted count matches the actual count with $Pa \in [0, 1]$.

We also introduce the metric $Pa \pm 1$ that allows a mismatch of one leaf between prediction and ground truth. For false positive detections, we use a ground truth leaf count of zero.

As a baseline, we provide comparisons with the popular state-of-the-art Mask R-CNN [12] framework for the task of object instance segmentation in two stages. We adopt this approach to our task, such that it is most comparable to our proposed method and can perform the same predictions.

First, a set of reference boxes with predefined sizes and aspect ratios serves as detection candidates. In the second stage, each positive box is refined and assigned to a set of categories. Simultaneously, Mask R-CNN predicts a $28 \times 28\,\text{px}$ binary semantic mask for each object.

For a comparison with our approach, we need to extract the total leaf count per plant from each binary mask. Thus, we encode each leaf keypoint in an object's ground truth mask, which allows us to estimate its total leaf count at inference by applying a simple post-processing procedure. This representation of leaves is commonly used for this task [30], [15] and is similar to our heatmap of leaf keypoints.

For training, we create a binary mask for each plant where its ground truth leaf keypoints are encoded with a fixed

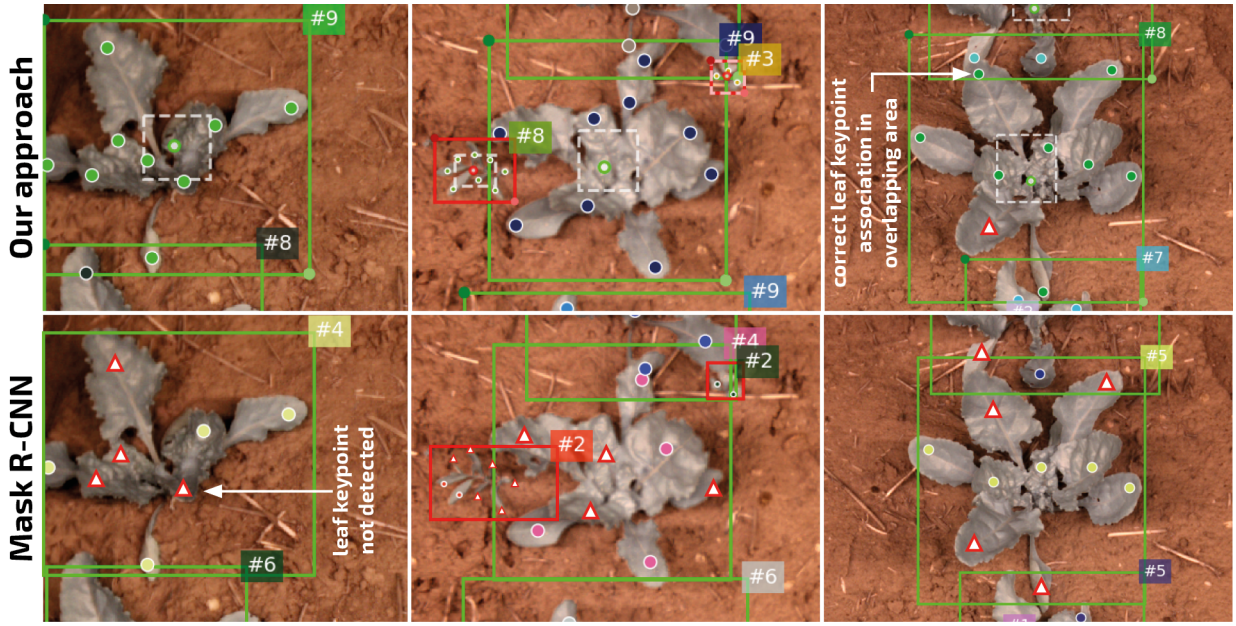

Fig. 5. Qualitative results of our approach (top row) and Mask R-CNN (bottom row). We show magnified views of our results to improve visibility. We predict the location of crops (green) and weeds (red) by their bounding boxes. Leaf keypoints associated with the same plant share the same color as well as its related total leaf count (#). Left: Our approach achieves a more accurate leaf count estimation for crops. Leaves that are not detected are marked with red triangles. Middle: We observe the same effect for weeds, which are characterized by a more heterogeneous visual appearance. Right: Our approach is more robust in complex scenes where bounding boxes of different instances overlap and associates leaf keypoints correctly. (Best viewed in color).

radius r_{LK}. We choose $r_{\text{LK}} = 2\,\text{px}$ such that nearby leaves do not intersect to avoid an under-counting of the total leaf count per plant. Thus, the binary mask of each plant is defined analogously to our heatmap of leaf keypoints and is well suited for the task of counting. We define the mask loss as a variant of focal loss, see Eq. (1).

For inference, we resize each predicted mask and binarize it at a confidence threshold of 0.5. Finally, we apply a dilation to merge fragmented leaf keypoints and extract the total leaf count per plant by computing its connected components.

B. Training and Testing Details

During the training of our approach, we use data augmentation to reduce overfitting by horizontally and vertically flipping the original image. The radius r_{LK} of 2D Gaussians is set to $2\,\text{px}$. We define a weighted multi-task loss function:

$$\mathcal{L} = \alpha_{\text{det}}\mathcal{L}_{\text{det}} + \beta_{\text{cluster}}\mathcal{L}_{\text{cluster}} + \gamma_{\text{off}}\mathcal{L}_{\text{off}}, \tag{6}$$

$$\mathcal{L}_{\text{det}} = \mathcal{L}^{\text{co}}_{\text{det}} + \mathcal{L}^{\text{ce}}_{\text{det}} + \mathcal{L}^{\text{leaves}}_{\text{det}}, \tag{7}$$

$$\mathcal{L}_{\text{cluster}} = \mathcal{L}^{\text{co}}_{\text{pull}} + \mathcal{L}^{\text{co}}_{\text{push}} + \mathcal{L}^{\text{leaves}}_{\text{pull}} + \mathcal{L}^{\text{leaves}}_{\text{push}}, \tag{8}$$

$$\mathcal{L}_{\text{off}} = \mathcal{L}^{\text{co}}_{\text{off}} + \mathcal{L}^{\text{ce}}_{\text{off}}, \tag{9}$$

composed of the focal losses in $\mathcal{L}_{\text{det}}$ to detect corner (co), center (ce) and leaf keypoints of all objects as well as the clustering losses $\mathcal{L}_{\text{cluster}}$ with the objective of small distances between associative embeddings. We set the $L1$-distance threshold of non-associative embeddings in feature space to $\Delta = 1$ in all experiments. The smooth $L1$ losses in $\mathcal{L}_{\text{off}}$ train the network to adjust coarse locations of corner and center predictions due to down-sampled heatmaps. We set the weights of the composed loss function $\mathcal{L}$ to $\alpha_{\text{det}} = 1$, $\beta_{\text{cluster}} = 0.25$, and $\gamma_{\text{off}} = 1$. The dimension D of our embedding vectors is set to 16 in all experiments. For details of corner and centers losses, we refer to Law *et al.* [19]. To optimize the multi-task loss, we use Adam and set the learning rate to 5×10^{-4} with a epoch-wise decay of 0.995 and train for 512 epochs with a batch size of 2. We set weight decay to 1×10^{-4} and use dropout with $p = 0.1$ in the FPN.

For inference, we first apply NMS with $k_{\text{NMS}} = 3$ on all predicted heatmaps. Next, we select top 50 center points, top-left, and bottom-right corners as well as top 250 leaf keypoint predictions by confidence score, which we previously denoted as top-k and top-n respectively. We associate a pair of corners if the distance of their corresponding embeddings is less than $\theta_{\text{co}} = 0.95$, whereby each top-left corner has a maximum of $j_{\text{max}} = 1$ related bottom-right corners and vice versa. Additionally, we define scale-aware center regions for small, medium, and large boxes as proposed by Duan *et al.* [7]. Thus, the center region of a small box is identical to its bounding box while we assign tighter regions to medium and large boxes to reduce the rate of false positives. Finally, we keep only boxes that contain a center point within their center region and where for each keypoint of its triplet $\theta_{\text{score}} > 0.5$. To associate a leaf keypoint with $\theta_{\text{score}} > 0.5$ to its related instance, we compute the distances to all averaged top-left and bottom-right corner embeddings of valid bounding boxes and assign it to the nearest neighbor of the same category if the distance is less than $\theta_{\text{leaf}} = 0.95$.

We test our approach on a NVIDIA GTX 1080 and obtain an average inference time of $200\,\text{ms}$ per image and for Mask R-CNN $250\,\text{ms}$.

TABLE II
OBJECT DETECTION PERFORMANCE OF OUR APPROACH IN COMPARISON WITH STATE-OF-THE-ART METHOD MASK R-CNN.

Approach	AP	AP^{50}	AP^{75}	AP_S	AP_M	AP_L	AR_S	AR_M	AR_L
Ours (mean average)	**56.5**	72.5	**61.8**	**19.4**	**50.1**	**48.6**	27.4	**57.4**	**51.7**
Mask R-CNN (mean average)	47.2	**80.0**	48.4	17.6	35.6	43.7	**29.6**	46.9	50.8
Ours (crop)	**71.4**	82.0	**76.0**	5.2	**48.5**	**75.0**	12.9	**56.5**	**77.7**
Mask R-CNN (crop)	62.2	**92.0**	72.4	**8.1**	32.8	65.5	**20.8**	45.5	73.5
Ours (weed)	**41.5**	63.0	**47.6**	**33.7**	**51.7**	**22.2**	**41.5**	**58.3**	25.7
Mask R-CNN (weed)	32.2	**68.1**	24.3	27.2	38.5	21.8	38.3	48.3	**28.0**

TABLE III
DIFFERENCE IN LEAF COUNTING OF OUR APPROACH IN COMPARISON WITH STATE-OF-THE-ART METHOD MASK R-CNN. NOTATION: MEAN (STD.).

Approach	*DiC*	DiC_S	DiC_M	DiC_L	*aDiC*	$aDiC_S$	$aDiC_M$	$aDiC_L$
Ours (crop)	**0.20 (1.69)**	**-0.50 (0.63)**	-0.25 (0.56)	**0.26 (1.77)**	**1.16 (1.24)**	**0.64 (0.48)**	**0.33 (0.52)**	**1.26 (1.27)**
Mask R-CNN (crop)	2.26 (2.20)	-0.67 (0.67)	**-0.08 (0.83)**	2.63 (2.12)	2.54 (1.88)	0.67 (0.67)	0.57 (0.61)	2.84 (1.83)
Ours (weed)	**0.04 (1.64)**	**0.24 (1.17)**	**0.02 (1.28)**	**-0.48 (3.83)**	**0.99 (1.31)**	**0.89 (0.80)**	**0.82 (0.99)**	**2.59 (2.86)**
Mask R-CNN (weed)	0.19 (2.85)	-0.46 (1.42)	0.06 (1.34)	2.50 (6.82)	1.47 (2.45)	1.15 (0.96)	0.93 (0.97)	4.81 (5.44)

TABLE IV
PERCENTAGE OF AGREEMENT IN LEAF COUNTING OF OUR APPROACH IN COMPARISON WITH STATE-OF-THE-ART METHOD MASK R-CNN.

Approach	*Pa*	Pa_S	Pa_M	Pa_L	$Pa \pm 1$	$Pa_S \pm 1$	$Pa_M \pm 1$	$Pa_L \pm 1$
Ours (crop)	**0.33**	0.35	**0.69**	**0.29**	**0.70**	**1.00**	**0.97**	**0.67**
Ours (crop - heatmap only)	0.28	0.35	0.66	0.24	0.65	1.00	0.97	0.60
Mask R-CNN (crop)	0.13	**0.44**	0.48	0.07	0.35	0.88	0.96	0.26
Ours (weed)	**0.41**	**0.35**	**0.46**	**0.26**	**0.77**	**0.78**	**0.80**	**0.48**
Ours (weed - heatmap only)	0.41	0.35	0.46	0.26	0.77	0.78	0.80	0.48
Mask R-CNN (weed)	0.32	0.28	0.40	0.06	0.68	0.66	0.77	0.31

C. Performance on Crop-Weed Detection

The first experiment supports the claim that our approach accurately detects crops and weeds of different growth stages and is superior to state-of-the-art two-stage approaches. In Tab. II we show test set results for the evaluated approaches. The results show that we achieve a higher detection performance most importantly for value crops. We report a test AP of $71.4\,\%$ considering crops of all object area ranges, an improvement of 9.2 percent points over $62.2\,\%$ obtained by Mask-RCNN. Our approach suggests a lower rate of incorrect predictions across medium $(48.5\,\%)$, and large $(75.0\,\%)$ crops as well as an higher AR. This difference is critical for precision farming to predict yield potential. In terms of weed detection, we achieve an AP improvement from $32.2\,\%$ to $41.5\,\%$ and an improved AR across small and medium area ranges. This is highly relevant for weed treatment in the early growth stages.

D. Performance on Leaf Count Estimation

In this experiment, we show the capability of our approach to estimate leaf count for in-field phenotyping of crops and weeds, see Fig. 5. In Tab. III and Tab. IV we show that our method outperforms Mask R-CNN in this task for high value crops. We achieve a *Pa* of $33\,\%$ for crops across all area ranges, an improvement of 20 percent points over $13\,\%$ reported by Mask R-CNN. We also argue that a count offset of one leaf is within the accuracy of a manual inspection and evaluate our predictions based on this metric. The high increase of performance in $Pa \pm 1$ for small crops (from $35\,\%$ to $100\,\%$) relates to the fact that those plants have only a few leaves per se, see Tab. I. Our network also shows superior performance for later growth stages of medium $(97\,\%)$ and large crops $(67\,\%)$ in contrast to Mask R-CNN $(96\,\%$ or $26\,\%)$ allowing for a more accurate monitoring and determination of growth stages of individual plants. Those results are supported in terms of *aDiC* for high value crops achieving a smaller mean in *aDiC* of (1.16) with a higher precision (1.24) opposed to Mask R-CNN $(2.54$ or $1.88)$. Contrarily, different weed types are characterized by a more heterogeneous visual appearance with many leaves leading to a drop in performance, especially for large weeds. We still obtain a lower mean in *aDiC* of 2.59 for large weeds in comparison with Mask R-CNN (4.81) with higher precision.

E. Ablation Study

Estimating the leaf count based on associative embeddings is a key component of our approach to deal with overlapping leaves of different plants. To demonstrate its contribution, we provide a comparison to a leaf count method, which is only based on the leaf keypoint heatmap but with the same hyperparameters (Tab. IV). For each predicted bounding box we count the total number of leaf keypoints inside but reject keypoints of different categories (Fig. 4). The results of our approach show an improvement for crops across medium and large object area scales in terms of Pa_M and Pa_L demonstrating the importance of associative embeddings. Crops at those growth stages are characterized by a high IoU (Tab. I) and thus more likely to contain leaves of different plants within their bounding box. We see the highest increase of 7 percent points (from $60\,\%$ to $67\,\%$) in terms of $Pa \pm 1$ for large crops, which accordingly have the highest IoU. Furthermore, small crops and weeds

across all area scales are characterized by a lower IoU in our dataset (Tab. I) and thus are less likely to contain leaves of different plants within their bounding box. Their leaf count performance is not affected by using associative embeddings. Our results convey that using associative embeddings increases the leaf count performance of overlapping plants but does not affect the performance of plants with little or no overlap.

V. Conclusion

In this paper, we presented a novel approach to analyze plant development in complex scenes on agricultural fields. Our system uses camera images as inputs to predict the bounding boxes of crops and weeds as well as leaf keypoints associated with each plant to estimate its leaf count. These parameters are commonly used to describe vegetative stages. Thus, our approach provides plant breeders with the ability to assess phenotypic traits in an automated fashion more frequently, objectively, and on a large scale in comparison with conventional, manual methods. In contrast to related methods, which predict additional keypoints for each object, our approach does not rely on any a priori assumption about the number of keypoints associated with each object and is thus well suited for the task of per plant leaf count estimation. We implemented and thoroughly tested our approach on data from a real agricultural field. Our experiments show that our approach achieves superior performance with respect to a Mask R-CNN-based approach on complex scenes with overlapping plants in different growth stages.

References

[1] E.E. Aksoy, A. Abramov, F. Wörgötter, H. Scharr, A. Fischbach, and B. Dellen. Modeling leaf growth of rosette plants using infrared stereo image sequences. *Computers and Electronics in Agriculture*, 110:78–90, 2015.

[2] S. Bargoti and J.P. Underwood. Deep Fruit Detection in Orchards. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2017.

[3] D. Boyes, A. Zayed, R. Ascenzi, A.J. McCaskill, N.E. Hoffman, K. Davis, and J. Görlach. Growth stage–based phenotypic analysis of Arabidopsis: a model for high throughput functional genomics in plants. *The Plant Cell*, 13(7):1499–1510, 2001.

[4] BSA Bundessortenamt. Beschreibende Sortenliste Getreide, Mais, Öl- und Faserpflanzen, Leguminosen, Rüben, Zwischen-früchte, 2013.

[5] D. Chen, K. Neumann, S. Friedel, B. Kilian, M. Chen, T. Altmann, and C. Klukas. Dissecting the phenotypic components of crop plant growth and drought responses based on high-throughput image analysis. *The Plant Cell*, 26(12):4636–4655, 2014.

[6] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, June 2009.

[7] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian. CenterNet: Keypoint Triplets for Object Detection. In *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision (ICCV)*, 2019.

[8] M. Everingham, L. Van Gool, C.K. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *Intl. Journal of Computer Vision (IJCV)*, 88(2):303–338, 2010.

[9] F. Fiorani and U. Schurr. Future scenarios for plant phenotyping. *Annual review of plant biology*, 64:267–291, 2013.

[10] V.M. Giuffrida, P. Doerner, and S. Tsaftaris. Pheno-deep counter: a unified and versatile deep learning architecture for leaf counting. *The Plant Journal*, 96(4):880–890, 2018.

[11] F. Golbach, G. Kootstra, S. Damjanovic, G. Otten, and R. van de Zedde. Validation of plant part measurements using a 3d reconstruction method suitable for high-throughput seedling phenotyping. *Machine Vision and Applications*, 27(5):663–680, 2016.

[12] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *Proc. of the IEEE Intl. Conf. on Computer Vision (ICCV)*, 2017.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[14] M. Hess, G. Barralis, H. Bleiholder, L. Buhr, T.H. Eggers, H. Hack, and R. Stauss. Use of the extended BBCH scale—general for the descriptions of the growth stages of mono; and dicotyledonous weed species. *Weed Research*, 37(6):433–441, 1997.

[15] Y. Itzhaky, G. Farjon, F. Khoroshevsky, A. Shpigler, and A. Bar-Hillel. Leaf counting: Multiple scale regression and detection using deep CNNs. In *Proc. of British Machine Vision Conference (BMVC)*, 2018.

[16] A.J. Jhala. Guide for Weed, Disease, and Insect Management in Nebraska, 2016.

[17] R. Kieloch and K. Domaradzki. The role of the growth stage of weeds in their response to reduced herbicide doses. *Acta Agrobotanica*, 64(4):259–266, 2011.

[18] J.P. Kumar and S. Domnic. Image based leaf segmentation and counting in rosette plants. *Information Processing in Agriculture*, 6(2):233–246, 2019.

[19] H. Law and J. Deng. CornerNet: Detecting Objects as Paired Keypoints. In *Proc. of the Europ. Conf. on Computer Vision (ECCV)*, 2018.

[20] T.Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature Pyramid Networks for Object Detection. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[21] T.Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal Loss for Dense Object Detection. In *Proc. of the IEEE Intl. Conf. on Computer Vision (ICCV)*, 2017.

[22] J. Long, E. Shelhamer, and T. Darrell. Fully Convolutional Networks for Semantic Segmentation. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[23] P. Lottes, J. Behley, N. Chebrolu, A. Milioto, and C. Stachniss. Robust joint stem detection and crop-weed classification using image sequences for plant-specific treatment in precision farming. *Journal of Field Robotics (JFR)*, 37:20–34, 2020.

[24] P. Lottes, J. Behley, A. Milioto, and C. Stachniss. Fully convolutional networks with sequential information for robust crop and weed detection in precision farming. *IEEE Robotics and Automation Letters (RA-L)*, 3:3097–3104, 2018.

[25] P. Lottes, M. Hoeferlin, S. Sander, M. Müter, P. Schulze Lammers, and C. Stachniss. An effective classification system for separating sugar beets and weeds for precision farming applications. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2016.

[26] C.S. McCool, T. Perez, and B. Upcroft. Mixtures of Lightweight Deep Convolutional Neural Networks: Applied to Agricultural Robotics. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2017.

[27] A. Milioto, P. Lottes, and C. Stachniss. Real-time blob-wise sugar beets vs weeds classification for monitoring fields using convolutional neural networks. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2017.

[28] A.K. Mortensen, M. Dyrmann, H. Karstoft, R. N. Jörgensen, and R. Gislum. Semantic Segmentation of Mixed Crops using Deep Convolutional Neural Network. In *Proc. of the Intl. Conf. of Agricultural Engineering (CIGR)*, 2016.

[29] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proc. of the Advances in Neural Information Processing Systems (NIPS)*, 2015.

[30] H. Scharr, T.P. Pridmore, and S.A. Tsaftaris. Computer Vision Problems in Plant Phenotyping, CVPPP 2017–Introduction to the CVPPP 2017 Workshop Papers. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops*, 2017.

[31] W. Shi, R. van de Zedde, H. Jiang, and G. Kootstra. Plant-part segmentation using deep learning and multi-view vision. *Biosystems Engineering*, 187:81–95, 2019.

[32] M. Simić, V. Dragičević, S. Knežević, M. Radosavljević, Ž. Dolijanović, and M. Filipović. Effects of applied herbicides on crop productivity and on weed infestation in different growth stages of sunflower (helianthus annuus l.). *Helia*, 34(54):27–38, 2011.

[33] J.R. Teasdale and D.W. Shirley. Influence of Herbicide Application Timing on Corn Production in a Hairy Vetch Cover Crop. *Journal of Production Agriculture*, 11(1):121–125, 1998.

[34] J. Wang, J. He, Y. Han, C. Ouyang, and D. Li. An adaptive thresholding algorithm of field leaf image. *Computers and Electronics in Agriculture*, 96:23–39, 2013.

[35] X. Zhou, D. Wang, and P. Krähenbühl. Objects as Points. *arXiv preprint*, 2019.