In-Field Phenotyping Based on Crop Leaf and Plant Instance Segmentation

Jan Weyler[1] Federico Magistri[1] Peter Seitz[2] Jens Behley[1] Cyrill Stachniss[1]
[1]University of Bonn, Germany [2]Robert Bosch GmbH, Germany

Abstract

A detailed analysis of a plant's phenotype in real field conditions is critical for plant scientists and breeders to understand plant function. In contrast to traditional phenotyping performed manually, vision-based systems have the potential for an objective and automated assessment with high spatial and temporal resolution. One of such systems' objectives is to detect and segment individual leaves of each plant since this information correlates to the growth stage and provides phenotypic traits, such as leaf count, coverage, and size. In this paper, we propose a vision-based approach that performs instance segmentation of individual crop leaves and associates each with its corresponding crop plant in real fields. This enables us to compute relevant basic phenotypic traits on a per-plant level. We employ a convolutional neural network and operate directly on drone imagery. The network generates two different representations of the input image that we utilize to cluster individual crop leaf and plant instances. We propose a novel method to compute clustering regions based on our network's predictions that achieves high accuracy. Furthermore, we compare to other state-of-the-art approaches and show that our system achieves superior performance. The source code of our approach is available[1].

1. Introduction

Crop production is key for our society to provide feed, food, and other resources. During the growth of a plant, the development of its functional body is affected by a dynamic process between its genotype, the performed management, and the environment [4]. Thus, plant scientists and breeders continuously assess phenotypic traits as an expression of the genotype for individual plants to generate new genetic variations of crops that show desired traits. Outside greenhouses, this in-field assessment is conventionally done manually, which is time-consuming [20]. In contrast, vision-based systems have the prospect to perform this assessment at a large scale, in less time, and more objectively [30].

[1]https://github.com/PRBonn/leaf-plant-instance-segmentation

Figure 1: Our approach takes images of real fields (top) and provides an instance segmentation for individual crop leaves (middle) and plants (bottom), each represented by a particular color.

A key target of these systems is to predict the total number of leaves per plant. This information is commonly used to describe plant growth stages, which are linked to yield potential and plant performance [13]. However, when studying the plant growth in more detail, it is also essential to segment individual leaves in order to determine the leaf size and shape to get a clearer response [29]. At the same time, obtaining this refined information on a per-pixel level is challenging, particularly in uncontrolled in-field conditions with multiple plants. In this environment, each segmented leaf needs to be associated with a specific plant on the field to enable a reliable analysis on a per-plant level.

In this paper, we address the problem of automated, vision-based phenotyping to detect and segment individual leaves of crops based on images taken from real agricultural fields that we associate to specific crop plants to extract relevant basic phenotypic traits on a per-plant level. This provides plant scientists and breeders with reproducible phenotyping information with a high spatial and temporal resolution in contrast to manual field assessments.

The main contribution of this work is a vision-based approach that performs a simultaneous instance segmentation of individual crop leaves and plants, as shown in Fig. 1. We target sugar beets as crops. Our approach computes binary

segmentation masks for all crop leaves in the field and associates them with their corresponding plant. This enables us to compute relevant basic phenotypic traits for individual crops. Our method is a bottom-up approach based on an end-to-end trainable single-shot convolutional neural network (CNN). We generate two different representations of the input image that are eligible to cluster individual crop leaf and plant instances within a predicted clustering region.

We make the following four claims about our approach. First, our bottom-up approach accurately performs a simultaneous instance segmentation of individual crop leaves and plants on real agricultural fields based on a single-shot CNN. Second, this allows us to derive relevant basic phenotypic traits for individual crops in the field. Third, in both tasks, our approach is competitive with different state-of-the-art methods. Fourth, for the clustering of crop leaves and plant instances, we present a novel method to specify clustering regions by full covariance matrices predicted by our network that shows superior performance compared to previous methods.

2. Related Work

There has been significant progress towards vision-based methods for semantic and instance segmentation in real agricultural fields. However, most methods for image-based phenotyping have been applied in laboratory environments.

Semantic Segmentation. Most recent approaches use CNNs to perform semantic segmentation based on images of real fields and provide a pixel-wise classification. Lottes *et al*. [15] propose a crop-weed classification system based on sequential image data recorded by agricultural robots, which exploits the spatial arrangement of crops and weeds to perform robust pixel-wise labeling. McCool *et al*. [17] propose a method for crop-weed classification that learns lightweight CNN models, which are appropriate to run on robotic platforms and achieve high accuracy for the task of weed segmentation. Milioto *et al*. [19] perform semantic segmentation of crops based on RGB and near-infrared (NIR) images but also compute multiple vegetation indices in a preprocessing step to support the training. Unlike our approach, these methods do not detect leaf or plant instances, which is key to extract morphological plant traits.

Instance Segmentation. Contrary, recently proposed image-based instance segmentation methods aim at detecting and segmenting individual plants. Champ *et al*. [3] rely on Mask R-CNN [8] to perform instance segmentation for different crops and weeds on real fields based on RGB images to target weed control. In contrast, Milioto *et al*. [18] propose a vision-based, two-stage approach, which first detects single plants based on RGB and NIR information and feeds each to a CNN classifying whether it is a crop or weed. Opposite to these plant-based methods, Morris [21] performs detection and segmentation of overlapping leaves in dense foliage images based on a pyramid CNN, which detects and discriminates leaf boundaries from interior textures. In contrast to our approach, these methods exclusively detect and segment plant or leaf instances but not both simultaneously. Thus, these methods are incapable of extracting per-plant leaf count.

Phenotyping. Most methods extract morphological plant traits based on images or 3D models of plants acquired individually in the laboratory. Kulikov [12] presents an instance segmentation approach to detect leaves based on images of single plants captured in the laboratory. He proposes a two-stage method that first specifies target embeddings, which are subsequently learned by a CNN and allow for a clustering approach at inference time to recover each instance. In contrast, Shi *et al*. [30] rely on a multi-view approach that performs semantic and instance segmentation based on Mask R-CNN for multiple images of single tomato plants. They combine the predictions of different viewpoints to 3D point clouds and perform instance segmentation of leaves, stems, and nodes. Magistri *et al*. [16] aim at an automated tracking of phenotypic traits over time based on 3D models of individual growing plants. Itzhaky *et al*. [10] propose a CNN to generate a heatmap of leaf keypoints for images of single plants and feed this map to a non-linear regression model to predict the total number of leaves per plant. In contrast to these methods, our approach does not rely on images of single plants but is applied in real fields. Weyler *et al*. [32] jointly detect the bounding box of individual plants and per-plant leaf keypoints based on a single-shot detection approach in images of real fields to compute the total number of leaves per plant. However, this method does not segment individual leaves nor plants but provides coarse keypoints that are not suitable to determine leaf size and shape. In contrast, our approach obtains refined information on a per-pixel level instead of coarse leaf keypoints. This setting is challenging since images of real fields usually contain multiple plants. Thus, each segmented leaf needs to be associated with a specific plant on the field to compute relevant basic phenotypic traits on a per-plant level. To account for this association problem, our approach has some relations to work on human pose estimation [23]. However, they assume that the number of parts per instance is known a priori, which is not reasonable for plants in different growth stages.

3. Our Approach

The main objective of our approach is to generate a binary segmentation mask for each leaf of a crop and to associate it with a specific crop plant based on images of real agricultural fields. Thus, we perform a simultaneous instance segmentation of individual crop leaves and plants. Accordingly, we can determine the shape and size of individual leaves but also the number of leaves per crop, which

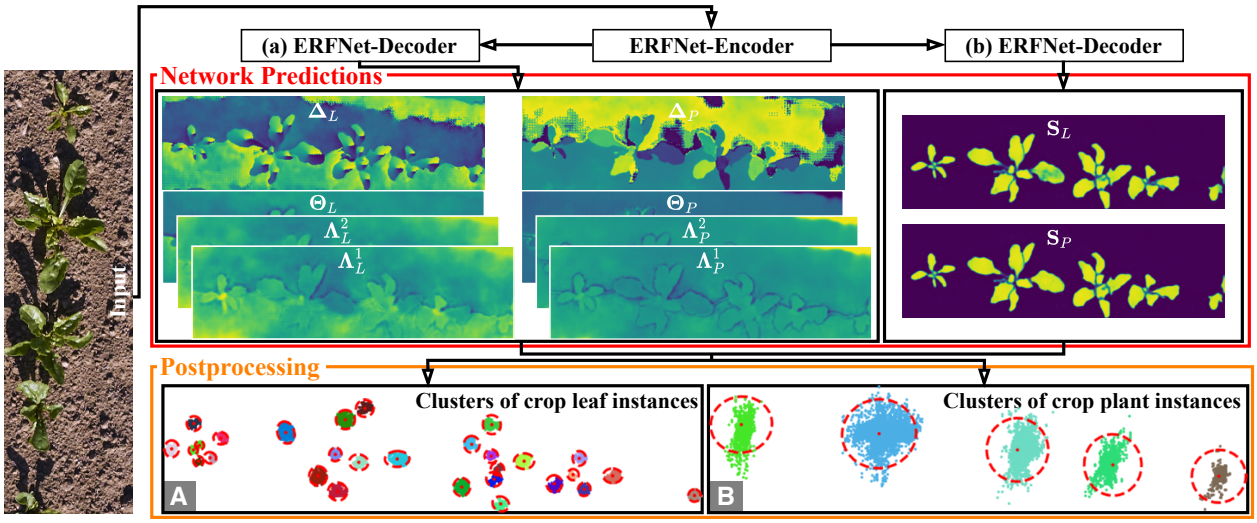

Figure 2: The network architecture of our approach. Based on RGB images, we predict offset maps $\mathbf{\Delta}_L$ and $\mathbf{\Delta}_P$ that translates each pixel of a crop leaf and plant into a clustering region around its associated center. The clustering regions are specified by covariance matrices. We compute the covariance matrices for crop leaf instances based on the predicted feature maps $\mathbf{\Theta}_L$, $\mathbf{\Lambda}_L^1$, and $\mathbf{\Lambda}_L^2$ and the covariance matrices for crop plant instance based on $\mathbf{\Theta}_P$, $\mathbf{\Lambda}_P^1$, and $\mathbf{\Lambda}_P^2$. Besides, we predict the feature map $\mathbf{S}_L$ and $\mathbf{S}_P$ to recover the centers of individual crop leaf and plant instances, respectively. We exploit our network's predictions to generate two different representations (A, B) of the input image in an automated postprocessing step that we utilize to cluster individual crop leaf and plant instances.

is highly relevant to perform phenotyping [20, 29].

To achieve this twofold instance segmentation, we propose a bottom-up approach based on a CNN whose architecture is described in Sec. 3.1. Our network takes an RGB image as input, which we feed into an encoder-decoder architecture based on ERFNet [26] to compute dense predictions. We split the decoder into two branches, which are labeled as (a) and (b) in Fig. 2. We design the first decoder (a) to predict offsets that enforce pixels of individual crop leaves to point into a leaf-specific region around the leaf center they belong to. Simultaneously, we predict another set of offsets that enforces pixels of individual crop leaves to point into a plant-specific region around the plant center they belong to (Sec. 3.2). In addition, this decoder predicts the parameters required to compute clustering regions around each center (Sec. 3.3). Based on the prediction of the second decoder (b), we predict the center locations of each instance (Sec. 3.4). Finally, we generate two different representations of the input image based on these predictions that we utilize to cluster each crop leaf and plant instance with an automated post-processing step (Sec. 3.5) applied after the CNN, as shown at the bottom of Fig. 2.

3.1. General Architectural Concept

Inspired by the recent success of bottom-up approaches for instance segmentation [2, 23], we design an enhanced version of the method proposed by Neven *et al*. [22] that enables a simultaneous instance segmentation of individual crop leaves and their corresponding plant. We explicitly model the instance segmentation of a crop plant as the union of the binary masks of its associated leaves. The original method [22] does not allow to model a simultaneous instance segmentation and is also more restricted in the design of clustering regions (Sec. 3.3).

The objective of our proposed twofold instance segmentation is to cluster a set of 2D pixel coordinates $X = \{0, 1, ..., W-1\} \times \{0, 1, ..., H-1\}$ into a set of crop leaf instances $L = \{L_0, L_1, ..., L_{K-1}\}$ and crop plant instances $P = \{P_0, P_1, ..., P_{J-1}\}$, where $L_k \subset X$ and $P_j \subset X$. Let W and H denote the image width and height, respectively. Since, by nature, each leaf L_k is associated with a specific plant P_j on the field, we argue that each plant instance is defined as the union of its associated leaves.

To achieve the desired clustering, we learn two offset vectors $\Delta\mathbf{l}_i$ and $\Delta\mathbf{p}_i$ for each pixel $\mathbf{x}_i = (x_i, y_i) \in X$ such that the resulting spatial embeddings $\mathbf{l}_i = \mathbf{x}_i + \Delta\mathbf{l}_i$ and $\mathbf{p}_i = \mathbf{l}_i + \Delta\mathbf{p}_i$ (Fig. 3) point into a clustering region around the corresponding crop leaf center $\mathbf{C}_{L_k}$ and crop plant center $\mathbf{C}_{P_j}$, respectively. The centers correspond to the centroids of the k^{th} leaf or j^{th} plant.

Note that the spatial embedding $\mathbf{p}_i$ depends on the leaf embedding $\mathbf{l}_i$. Thus, to cluster an individual crop plant, we first translate each corresponding pixel to the center of its associated leaf and next to the center of its associated plant, see Fig. 3. Underlying this is that we consider sugar beet leaves to be easier to cluster due to their blob-like shape.

To perform the clustering, we propose two Gaussian functions $\phi_{L_k}(\cdot)$ and $\phi_{P_j}(\cdot)$ for each crop leaf L_k and plant P_j, which convert the distance between the embeddings $\mathbf{l}_i$ or $\mathbf{p}_i$ to their corresponding center $\mathbf{C}_{L_k}$ or $\mathbf{C}_{P_j}$ into a score

of belonging to that instance as:

$$\phi_{L_k}(\mathbf{l}_i) = \exp\left(-\frac{1}{2}(\mathbf{l}_i - \mathbf{C}_{L_k})^\top \mathbf{\Sigma}_{L_k}^{-1}(\mathbf{l}_i - \mathbf{C}_{L_k})\right), \quad (1)$$

$$\phi_{P_j}(\mathbf{p}_i) = \exp\left(-\frac{1}{2}(\mathbf{p}_i - \mathbf{C}_{P_j})^\top \mathbf{\Sigma}_{P_j}^{-1}(\mathbf{p}_i - \mathbf{C}_{P_j})\right), \quad (2)$$

where $\phi_{L_k}(\mathbf{l}_i) \in [0,1]$ and $\phi_{P_j}(\mathbf{p}_i) \in [0,1]$. A high score indicates that the embedding $\mathbf{l}_i$ or $\mathbf{p}_i$ is associated with the k^{th} crop leaf L_k or j^{th} plant instance P_j accordingly. In contrast, a low score indicates that this embedding is associated with a background pixel or another instance.

Note that we specify for each crop leaf and plant instance a specific covariance matrix $\mathbf{\Sigma}_{L_k} \in \mathbb{R}^{2\times2}$ and $\mathbf{\Sigma}_{P_j} \in \mathbb{R}^{2\times2}$, which determines the clustering region around the corresponding center for which spatial embeddings are considered to be part of the instance. These covariance matrices are learned in addition to the spatial embeddings by our proposed CNN and give the network the capability to adopt the clustering region around an object's center to its shape and orientation. This accounts for the nature of leaves that have a relatively blob-like shape in a variety of orientations.

During training, we optimize the intersection over union (IoU) between the predicted and the ground truth mask by feeding them to the Lovász Hinge loss [1]:

$$\mathcal{L}_{\text{leaves}} = \frac{1}{K}\sum_{k=0}^{K-1} \text{Lovász}\left(F_{L_k}, y^*_{L_k}\right), \quad (3)$$

$$\mathcal{L}_{\text{plants}} = \frac{1}{J}\sum_{j=0}^{J-1} \text{Lovász}\left(F_{P_j}, y^*_{P_j}\right), \quad (4)$$

where $y^*_{L_k} \in \{-1,1\}^{H\times W}$ and $y^*_{P_j} \in \{-1,1\}^{H\times W}$ denote the binary ground truth mask for each crop leaf and plant, respectively. Let $F_{L_k} \in \mathbb{R}^{H\times W}$ be the output scores of the model for the k^{th} crop leaf and $F_{P_j} \in \mathbb{R}^{H\times W}$ the output scores of for the j^{th} crop plant defined as:

$$F_{L_k}[y_i, x_i] = 2\phi_{L_k}(\mathbf{x}_i + \Delta\mathbf{l}_i) - 1 \qquad \forall \mathbf{x}_i \in X, \quad (5)$$

$$F_{P_j}[y_i, x_i] = 2\phi_{P_j}(\mathbf{x}_i + \Delta\mathbf{l}_i + \Delta\mathbf{p}_i) - 1 \;\forall \mathbf{x}_i \in X. \quad (6)$$

Here, we transform the scores to the range $[-1,1]$ such that the predicted binary masks $\hat{y}_{L_k}$ and $\hat{y}_{P_j}$ can be efficiently obtained by $\hat{y}_{L_k} = \text{sign}(F_{L_k})$ and $\hat{y}_{P_j} = \text{sign}(F_{P_j})$. This follows the definition of the Lovász Hinge loss [1] and sets the score threshold effectively to 0.5.

Based on Eq. (1) and Eq. (2) the network has multiple options to optimize the IoU between the predicted and the ground truth mask. First, the network can translate the pixel embeddings close to the desired centers and predict a small clustering region around an object's center specified by its covariance matrix. Second, the network can adapt the covariance matrix to the object's shape and orientation and predict minor translations for the spatial embeddings.

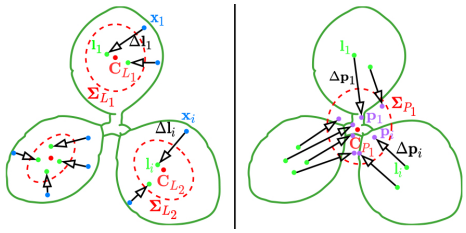

Figure 3: Clustering approach to perform instance segmentation for individual crop leaves (left) and plants (right). Our network predicts all entities to enforce pixels $\mathbf{x}_i$ to point into a clustering region (specified by covariance matrices $\mathbf{\Sigma}_{L_k}$ and $\mathbf{\Sigma}_{P_j}$) around each crop leaf $\mathbf{C}_{L_k}$ and plant center $\mathbf{C}_{P_j}$ and to perform the clustering. Note that we sample only a few pixels for visualization.

3.2. Spatial Embeddings

To translate individual pixels $\mathbf{x}_i$ towards their associated crop leaf and plant center, we apply the previously mentioned 2D offset vectors $\Delta\mathbf{l}_i$ and $\Delta\mathbf{p}_i$, respectively. Thus, our network predicts two offset maps denoted as $\mathbf{\Delta}_L \in \mathbb{R}^{2\times H\times W}$ and $\mathbf{\Delta}_P \in \mathbb{R}^{2\times H\times W}$. The channels contain the predicted offsets in x- and y-direction for all pixels. In Fig. 2, we show the orientation of these offsets encoded in a color scheme, *e.g.*, the offsets in $\mathbf{\Delta}_L$ point towards leaf centers and in $\mathbf{\Delta}_P$ towards plant centers.

Since in our case $W = 1024\,\text{px}$ and $H = 512\,\text{px}$, we generate a pixel coordinate map [22] $\mathbf{M}_{\text{coord}} \in \mathbb{R}^{2\times H\times W}$ that scales the x-coordinates of all pixels into the range $[0,2]$ and the y-coordinates into the range $[0,1]$. We apply a $\tanh(\cdot)$ activation function to the predicted offset maps to restrict its values to the range $[-1,1]$.

First, we compute $\mathbf{M}_{\text{coord}} + \mathbf{\Delta}_L$ to generate a representation of the image where all pixels belonging to a crop leaf are translated towards its associated center. Second, we compute $\mathbf{M}_{\text{coord}} + \mathbf{\Delta}_L + \mathbf{\Delta}_P$ to generate another representation where all pixels belonging to a crop plant are translated towards its associated center, as shown in Fig. 3. Note that during training, we do not compute gradients for $\mathbf{\Delta}_L$ but only for $\mathbf{\Delta}_P$ in the second step.

3.3. Instance Covariance Matrices

Our clustering functions described in Eq. (1) and Eq. (2) define each a clustering region around the instance centers determined by the associated covariance matrix. Thus, we propose a network architecture, which enables us to compute valid covariance matrices for each instance. In contrast, the original method proposed by Neven *et al.* [22] is restricted to predict diagonal covariance matrices and thus limited in the representation of clustering regions.

By definition, a valid covariance matrix needs to be symmetric, positive semi-definite, and square [6]. We must ensure that these properties hold for the predictions of our network. Thus, we exploit the properties of the spectral theory in linear algebra [7], which states that a symmet-

ric matrix $\mathbf{\Sigma} \in \mathbb{R}^{n \times n}$ can be decomposed as $\mathbf{\Sigma} = \mathbf{R}\mathbf{D}\mathbf{R}^\top$ or $\mathbf{\Sigma}^{-1} = \mathbf{R}\mathbf{D}^{-1}\mathbf{R}^\top$. Let $\mathbf{R} \in \mathbb{R}^{n \times n}$ be an orthogonal matrix, which contains the normalized eigenvectors stacked as columns and $\mathbf{D} \in \mathbb{R}^{n \times n}$ be a diagonal matrix containing the eigenvalues of $\mathbf{\Sigma}$. Since covariance matrices are positive semi-definite, all eigenvalues need to be non-negative [31].

First, we define $\mathbf{R} \in \mathbb{R}^{2\times 2}$ as a 2D rotation matrix $\mathbf{R}(\theta)$ determined by the angle θ. Since $\mathbf{R}(\theta) \in SO(2)$ is an orthogonal matrix [6] it meets the constraint mentioned above.

Second, we determine the diagonal matrix $\mathbf{D} \in \mathbb{R}^{2\times 2}$ by the two eigenvalues λ_1 and λ_2 as follows:

$$\mathbf{D}(\lambda_1, \lambda_2) = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \rightarrow \mathbf{D}^{-1}(\lambda_1, \lambda_2) = \begin{pmatrix} \lambda_1^{-1} & 0 \\ 0 & \lambda_2^{-1} \end{pmatrix}, \quad (7)$$

with the constraint that λ_1 and λ_2 are non-negative to account for positive semi-definiteness. Thus, a valid covariance matrix is determined by three values for our 2D case.

Note that the eigenvalues and eigenvectors of $\mathbf{\Sigma}$ completely determine the shape of our clustering region. In the case of $\theta = 0$ and $\lambda_1 = \lambda_2$, the region's shape is circular. In contrast it is elliptical but axis-aligned if $\lambda_1 \neq \lambda_2$. If in addition $\theta \neq 0$, it is rotated w.r.t. the axis as well.

Accordingly, we design our network to predict three values at each pixel location $\mathbf{x}_i$ to compute the covariance matrices $\mathbf{\Sigma}_{L_k}$ or $\mathbf{\Sigma}_{P_j}$ for each crop leaf or plant. However, we directly predict the inverse matrix $\mathbf{\Sigma}_{L_k}^{-1}$ or $\mathbf{\Sigma}_{P_j}^{-1}$ since these are required in Eq. (1) and Eq. (2).

Our network predicts three feature maps denoted as $\mathbf{\Theta}_L$, $\mathbf{\Lambda}_L^1$, and $\mathbf{\Lambda}_L^2$ which are $\in \mathbb{R}^{H\times W}$. The feature map $\mathbf{\Theta}_L$ predicts the angles θ_i, $\mathbf{\Lambda}_L^1$ predicts the first set of inverse eigenvalues $\lambda_{1,i}^{-1}$, and $\mathbf{\Lambda}_L^2$ predicts the second set of inverse eigenvalues $\lambda_{2,i}^{-1}$ for each pixel. We apply an exponential activation function to the feature maps $\mathbf{\Lambda}_L^1$ and $\mathbf{\Lambda}_L^2$ to enforce non-negative values to account for positive semi-definiteness. Besides, we multiply $\mathbf{\Theta}_L$ by $\frac{\pi}{2}$ to encourage the network to predict appropriate angles. We show these maps in a color-encoded representation in Fig. 2.

For training, we exploit the ground truth masks to set the parameters θ_{L_k}, λ_{1,L_k}^{-1}, and λ_{2,L_k}^{-1} of a crop leaf L_k to the average of all predictions belonging to this instance:

$$\theta_{L_k} = \sum_{\theta_i \in L_k} \frac{\theta_i}{|L_k|}, \; \lambda_{1,L_k}^{-1} = \sum_{\lambda_{1,i}^{-1} \in L_k} \frac{\lambda_{1,i}^{-1}}{|L_k|}, \; \lambda_{2,L_k}^{-1} = \sum_{\lambda_{2,i}^{-1} \in L_k} \frac{\lambda_{2,i}^{-1}}{|L_k|}. \quad (8)$$

At inference, we predict an instance center's location (Sec. 3.4) and at the same location we extract the three values from the associated feature maps.

Finally, we compute the inverse covariance matrix of the k^{th} crop leaf instance for the Gaussian in Eq. (1) as follows:

$$\mathbf{\Sigma}_{L_k}^{-1} = \mathbf{R}(\theta_{L_k})\, \mathbf{D}^{-1}(\lambda_{1,L_k}, \lambda_{2,L_k})\, \mathbf{R}^\top(\theta_{L_k}). \quad (9)$$

To compute the inverse covariance matrix $\mathbf{\Sigma}_{P_j}^{-1}$ of a crop plant instance P_j, we predict three additional feature maps $\mathbf{\Theta}_P$, $\mathbf{\Lambda}_P^1$, and $\mathbf{\Lambda}_P^2$ and follow the same procedure.

3.4. Instance Centers

During training, we compute the centers of each crop leaf and plant based on the ground truth masks. However, at inference time, we need to recover these centers to perform the clustering based on Eq. (1) and Eq. (2). Since the loss functions described in Eq. (3) and Eq. (4) enforce the spatial pixel embeddings $\mathbf{l}_i$ and $\mathbf{p}_i$ to lie close to their associated instance center, we need to sample an appropriate embedding for each crop leaf and plant and set them as their corresponding instance centers at inference time to perform the clustering. By appropriate embeddings, we refer to spatial embeddings which have a high score under the Gaussian function $\phi_{L_k}(\cdot)$ or $\phi_{P_j}(\cdot)$, since these are close to the ground truth center by definition of Eq. (1) and Eq. (2).

During training, we generate a score map for each crop leaf and plant instance by passing all $\mathbf{l}_i$ and $\mathbf{p}_i$ to the corresponding function $\phi_{L_k}(\cdot)$ and $\phi_{P_j}(\cdot)$, respectively. We exploit these computations to train our network to predict two score maps that imitate these maps and thus are suitable to recover the centers of all crop leaves and plants. In the following, we denote these map as $\mathbf{S}_L \in \mathbb{R}^{H\times W}$ and $\mathbf{S}_P \in \mathbb{R}^{H\times W}$, as illustrated in Fig. 2. The map $\mathbf{S}_L$ should be equal to the score map computed for all crop leaf instances during training. Thus, it contains values close to 0 for all pixels whose associated embedding $\mathbf{l}_i$ belongs to the background and values close to 1 if the corresponding embedding lies close to a crop leaf center. The same holds for the map $\mathbf{S}_P$ but in contrast for all crop plant instances and their associated embeddings $\mathbf{p}_i$. We achieve this objective by the following regression loss functions [22]:

$$\mathcal{L}_{\text{C}_L} = \frac{1}{N} \sum_{i=0}^{N-1} \begin{cases} w\left(s_{L,i} - \phi_{L_k}(\mathbf{l}_i)\right)^2, & \text{if } s_{L,i} \in L_k \\ s_{L,i}^2, & \text{otherwise} \end{cases} \quad (10)$$

$$\mathcal{L}_{\text{C}_P} = \frac{1}{N} \sum_{i=0}^{N-1} \begin{cases} w\left(s_{P,i} - \phi_{P_j}(\mathbf{p}_i)\right)^2, & \text{if } s_{P,i} \in P_j \\ s_{P,i}^2, & \text{otherwise} \end{cases} \quad (11)$$

where $s_{L,i}$ defines the network's output of the previously defined map $\mathbf{S}_L$ for the i^{th} pixel, N is the total number of pixels, and w is a weight factor set to 10 in all experiments. The upper term in Eq. (10) regresses the i^{th} output of $\mathbf{S}_L$ to the score of the Gaussian function for the k^{th} crop leaf instance if and only if the i^{th} pixel belongs to this instance. Otherwise, we regress it to 0, as in that case the i^{th} pixel belongs to the background. The same applies to Eq. (11) but in this regard we consider crop plant instances. We apply a sigmoid activation function to the map $\mathbf{S}_L$ and $\mathbf{S}_P$ such that their values are in $[0, 1]$. During training, we compute the gradients only for $s_{L,i}$ and $s_{P,i}$ [22]. In Sec. 3.5, we provide more details about how to recover instance centers.

3.5. Postprocessing

At inference, we employ an automated clustering approach based on our network's predictions to perform instance segmentation. First, we cluster all crop leaves and subsequently merge these into clusters of individual crop plants. Thus, we consider a plant as the union of its leaves.

First, to predict the semantic mask of each crop leaf, we compute their spatial embeddings by $\mathbf{M}_{\text{coord}} + \mathbf{\Delta}_L$ but do not consider pixels which have a score ≤ 0.5 in the predicted map $\mathbf{S}_L$ since we judge them as background. Subsequently, we sample the pixel with the highest score in $\mathbf{S}_L$ and set the location of its associated embedding $\mathbf{l}_i$ as the center $\mathbf{C}_{L_1}$ of the first leaf instance L_1. This is in accordance with Eq. (1). The confidence score of L_1 is equal to the score extracted from $\mathbf{S}_L$. At the same location we extract the predicted angle θ_{L_1} and both inverse eigenvalues λ_{1,L_1}^{-1} and λ_{2,L_1}^{-1} from $\mathbf{\Theta}_L$, $\mathbf{\Lambda}_L^1$, and $\mathbf{\Lambda}_L^2$ to compute $\mathbf{\Sigma}_{L_1}^{-1}$ according to Eq. (9). We use these entities to compute the Gaussian in Eq. (1) for all spatial embeddings $\mathbf{l}_i$ and assign each pixel to L_1 if and only if the score $\phi_{L_1}(\mathbf{l}_i) > 0.5$. Then, we mask out all pixels assigned to this instance and do not consider them for clustering of other leaves to avoid multiple assignments. We repeat this process until all pixels in $\mathbf{S}_L$ with a score $s_{L,i} > 0.5$ are consumed. Thus, we do not need to specify the number of clusters explicitly.

Second, to predict the semantic mask of each crop plant, we iterate over the set of previously detected leaves and compute their spatial embeddings $\mathbf{p}_i = \mathbf{l}_i + \Delta\mathbf{p}_i$, where $\Delta\mathbf{p}_i$ is extracted from the predicted offset map $\mathbf{\Delta}_P$. This translates all previously computed crop leaf clusters towards the center of their associated crop plant, as shown on the right side of Fig. 3. Subsequently, we select the pixel with the highest score in $\mathbf{S}_P$ and set the location of its associated embedding $\mathbf{p}_i$ as the center $\mathbf{C}_{P_1}$ of the first crop plant P_1. This is in accordance with Eq. (2). The confidence score of P_1 is equal to the score extracted from $\mathbf{S}_P$. We compute $\mathbf{\Sigma}_{P_1}^{-1}$ in the same way as we did for leaves but based on $\mathbf{\Theta}_P$, $\mathbf{\Lambda}_P^1$, and $\mathbf{\Lambda}_P^2$. Finally, we compute the Gaussian in Eq. (2) for all embeddings $\mathbf{p}_i$ associated with a leaf L_k and assign a leaf to the plant P_1 if and only if for more than $50\,\%$ of its embeddings $\phi_{P_1}(\mathbf{p}_i) > 0.5$ holds true. Thus, we associate a crop leaf with a specific crop plant if the majority of its pixels point into the clustering region of this plant. As before with the leaves, we mask out pixels assigned to this plant and do not consider them in the further procedure. We repeat this process until all pixels in $\mathbf{S}_P$ with $s_{P,i} > 0.5$ are consumed, or all leaves are associated with a plant. Consequently, crop leaves are only associated with a single plant.

Finally, we obtain two image representations, shown at the bottom of Fig. 2. These allow for an instance segmentation of all crop leaves and plants (Fig. 1). Simultaneously, this operation associates each leaf with a specific crop plant and enables us to compute relevant basic phenotypic traits.

4. Experimental Evaluation

We present our experiments to show the capabilities of our approach and to support our key claims, which are: Our bottom-up approach (i) performs a simultaneous instance segmentation of individual crop leaves and plants on real agricultural fields, (ii) allows to compute relevant basic phenotypic traits, (iii) is competitive w.r.t. to state-of-the-art approaches, and (iv) our design decision to use full covariance matrices to specify clustering regions shows superior performance in contrast to related work.

Implementation Details. In all experiments, we train our network for 512 epochs using Adam optimizer [11] with a learning rate of $1 \cdot 10^{-3}$ and a polynomial learning rate decay $(1 - \frac{\text{epoch}}{\text{max\_epoch}})^{0.9}$. We define a multi-task loss as sum of Eq. (3), Eq. (4), Eq. (10), and Eq. (11).

Datasets. We evaluate our method on RGB images of sugar beet fields. The dataset contains 1316 images with a size of $1024\,\text{px} \times 512\,\text{px}$ and a ground sampling distance of $1\,\frac{\text{mm}}{\text{px}}$. The images are recorded with an unmanned aerial vehicle (UAV) equipped with a PhaseOne iXM-100 camera mounted in nadir view. We captured the images in real fields in uncontrolled conditions that cause shadows and variable illumination, as shown in Fig. 1. Thus, this data is more challenging compared to images captured in the laboratory [10, 12]. For training, we use $60\,\%$ of the entire dataset and $20\,\%$ to validate the hyperparameters. To evaluate the final metrics, we rely only on the remaining $20\,\%$.

In addition, we evaluate our method on the small but demanding CVPPP Leaf Segmentation Challenge (LSC) [28] as a popular benchmark. We follow best practice and use the sequence A1 with the highest number of baselines.

Evaluation Metrics. To evaluate the performance of our approach and to compare it with state-of-the-art methods, we calculate the average precision (*AP*) and average recall (*AR*) that are commonly used for instance segmentation [5]. We provide these metrics separately for crop leaves and plants since our approach computes a simultaneous instance segmentation for both. We differ between instances with an area scale $a < 1024\,\text{px}^2$ and $a \geq 1024\,\text{px}^2$ to account for different object sizes denoted as AP_S and AP_M.

Besides, we adopt the evaluation metrics commonly used for leaf segmentation in phenotyping [29]. We evaluate the *Absolute Difference in Count* (|*DiC*|) to measure the leaf count performance between the predicted and ground truth number of leaves. In contrast, *Percentage Agreement (Pa)* is the number of times the predicted leaf count matches the ground truth. The *Symmetric Best Dice (SBD)* measures the leaf segmentation accuracy by the average overlap between the predicted and ground truth mask for all leaves. In contrast, the *Foreground-Background Dice (FBD)* measures the plant segmentation accuracy. The values of *Pa*, *SBD*, and *FBD* values are $\in [0, 1]$, where higher values indicate more accurate predictions. For more details we refer to [29].

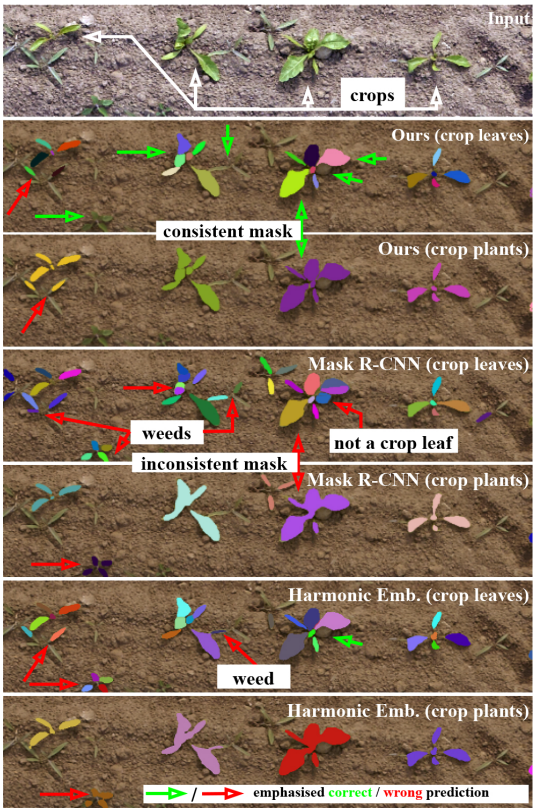

Figure 4: Qualitative results of our approach and both baselines. Note that we show cropped images and do not show the predicted bounding boxes of Mask R-CNN for reasons of clarity.

The performance of competing methods on the LSC is commonly specified in terms of *SBD* and $|DiC|$ [2, 12].

4.1. Comparison with the State of the Art

The first experiments evaluate the performance of our approach in comparison with other state-of-the-art methods.

First, we show that our approach is superior in comparison with Mask R-CNN [8, 34], a two-stage top-down method for instance segmentation. For comparison, we use models pre-trained on the COCO dataset [14] that leverage a ResNet50 model [9] and fine-tune it to our task. To provide a fair comparison, we train two networks based on Mask R-CNN. We train the first network with the objective to detect and segment all crop leaf instances and the second network to detect and segment all crop plants. Thus, both networks are experts for the specific task. However, we emphasize that our method performs both tasks at once. In Tab. 2, we show the results in terms of phenotypic metrics on the test set. Our proposed approach outperforms Mask R-CNN in all metrics. We achieve higher performance in terms of leaf count ($|DiC|$, *Pa*) per crop plant. In addition, the predicted masks for crop leaves of our method outperform the baseline by a wide margin in terms of *SBD*. Furthermore, our predicted masks for crop plants also have a

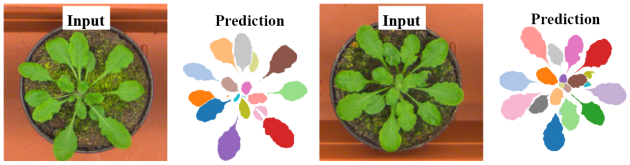

Figure 5: Qualitative results of our approach for the CVPPP LSC.

higher accuracy regarding *FBD*. In Fig. 4, we highlight that our predicted masks for crop leaves and plants are consistent since we explicitly model a crop plant as the union of its associated leaves (Sec. 3.5). In contrast, the predicted masks of Mask R-CNN are inconsistent. These results are supported in Tab. 1 in terms of *AP* and *AR* where our approach outperforms the baseline in most metrics. In Fig. 4, we show that our approach is less prone to confuse crop leaves and plants with leaves or plants of weeds which are commonly present on real agricultural fields.

Second, we show that our approach is competitive with another state-of-the-art method proposed by Kulikov [12] tailored to instance segmentation on biological images. This two-stage bottom-up method achieves state-of-the-art results on the popular CVPPP LSC. Similiar to Mask R-CNN it does not allow to perform a simultaneous instance segmentation of crop leaves and plants. Thus, we train two expert networks for each task. We denote this method as Harmonic Embeddings. In Tab. 2, we show that the performance in terms of predicted masks for crop leaves (*SBD*) and crop plants (*FBD*) only varies marginally in comparison with our method. We support these results visually in Fig. 4 and show that segmented crop leaves and plants differ only slightly. However, our approach achieves a higher performance in terms of leaf count per crop plant w.r.t. $|DiC|$ and *Pa*. Note that the method of Kulikov [12] does not predict confidence scores for object instances and thus does not support an evaluation in terms of *AP* and *AR*.

4.2. Performance on CVPPP LSC

The next experiments are designed to show that our approach achieves high performance on a popular leaf instance segmentation benchmark [28], see Fig. 5. In Tab. 3, we show that the performance of our method is on par with competing algorithms. Concerning the *SBD* metric, only the approach proposed by Wu *et al*. [33] achieves higher performance. However, their results rely on the ground-truth foreground masks, which we do not use.

We note that this competition addresses a less complex problem than our dataset since each image contains only a single plant. All competing methods are restricted to this assumption. In contrast, our network is also applicable to images of real fields that contain an arbitrary number of crops. The results convey that our approach covers a broader range of applications than competing methods but still achieves high performance in their targeted, restricted domain.

Table 1: Comparison of our method with Mask R-CNN based on average precision (*AP*) and average recall (*AR*) on our dataset.

Approach	*AP*	AP_{50}	AP_{75}	AP_{S}	AP_{M}	*AR*	AR_{S}	AR_{M}
Ours (crop leaves, $\boldsymbol{\Sigma}^{\mathrm{full}}$)	**48.7**	**82.5**	**54.6**	**46.8**	**78.2**	**57.3**	**55.6**	**81.4**
Ours (crop leaves, $\boldsymbol{\Sigma}^{\mathrm{diag}}$)	42.9	78.8	44.1	41.1	71.8	53.9	52.4	74.9
Mask R-CNN (crop leaves)	41.3	78.5	39.5	39.6	73.8	50.2	48.8	76.5
Ours (crop plants, $\boldsymbol{\Sigma}^{\mathrm{full}}$)	**60.4**	**93.8**	**73.5**	**28.1**	**63.7**	**68.0**	43.7	**71.1**
Ours (crop plants, $\boldsymbol{\Sigma}^{\mathrm{diag}}$)	56.5	93.1	67.2	25.6	60.2	65.6	43.8	68.4
Mask R-CNN (crop plants)	51.8	**93.8**	56.3	24.4	54.8	59.5	**46.3**	61.5
Ours (crop leaves, no $\boldsymbol{\Delta}$ + $\boldsymbol{\Sigma}^{\mathrm{full}}$)	**31.5**	**70.5**	**19.3**	**29.7**	**51.8**	**37.7**	**36.5**	**54.0**
Ours (crop leaves, no $\boldsymbol{\Delta}$ + $\boldsymbol{\Sigma}^{\mathrm{diag}}$)	24.3	68.7	5.7	22.9	41.7	31.4	30.6	42.8

Table 2: Evaluation of our dataset based on phenotypic metrics.

Approach	|*DiC*| (std.)↓	*Pa*↑	*SBD*↑	*FBD*↑
Ours ($\boldsymbol{\Sigma}^{\mathrm{full}}$)	**0.60** (**0.83**)	**0.55**	0.79	0.90
Ours ($\boldsymbol{\Sigma}^{\mathrm{diag}}$)	0.69 (0.94)	0.51	0.77	0.89
Mask R-CNN	1.53 (1.70)	0.30	0.68	0.86
Harmonic Emb.	0.68 (0.90)	0.51	**0.80**	**0.92**
Ours (no $\boldsymbol{\Delta}$ + $\boldsymbol{\Sigma}^{\mathrm{full}}$)	1.01 (**0.96**)	0.32	**0.66**	**0.78**
Ours (no $\boldsymbol{\Delta}$ + $\boldsymbol{\Sigma}^{\mathrm{diag}}$)	**1.00** (1.08)	**0.36**	0.63	0.73

Table 3: Evaluation on CVPPP LSC.

Approach	*SBD*↑	|*DiC*| (std.)↓
Recurrent IS + CRF [27]	66.6	1.1 (0.9)
IPK [24]	74.4	2.2 (1.3)
Discriminative Loss [2]	84.2	1.0 (-)
Recurrent with Attention [25]	84.9	**0.8** (1.0)
Harmonic Emb. [12]	89.9	3.0 (-)
W-Net [33]	**91.9**	-
Ours (crop leaves, $\boldsymbol{\Sigma}^{\mathrm{full}}$)	91.1	1.8 (2.4)

4.3. Ablation Studies

A key contribution of our method is the prediction of full covariance matrices based on the output of our CNN (Sec. 3.3) to compute clustering regions. This representation gives our network the capability to adjust the clustering region to an instance shape and orientation. To demonstrate its contribution, we train two different networks with the same hyperparameters. The former predicts full covariance matrices. For the latter, we remove the feature maps $\boldsymbol{\Theta}_L$ and $\boldsymbol{\Theta}_P$ and hence enforce diagonal covariance matrices $\boldsymbol{\Sigma}_{L_k}$ and $\boldsymbol{\Sigma}_{P_j}$ similar to Neven *et al.* [22]. Thus, we constrain axis-aligned clustering regions, which cannot adapt to an instance orientation. In Tab. 1 and Tab. 2 we show that the former network outperforms the latter in most metrics since it provides more degrees of freedom.

We also train two networks without the offsets $\boldsymbol{\Delta}_L$ and $\boldsymbol{\Delta}_P$. Hence, these networks have to adapt the clustering region to an instance shape and orientation to minimize the objectives in Eq. (3) and Eq. (4). We also predict full covariance matrices for the former network and diagonal covariance matrices for the latter. In the center of Fig. 6, we show that the former network effectively adjusts the clustering region to the orientation of each leaf and thus outperforms the latter network in most metrics, see Tab. 1 and Tab. 2.

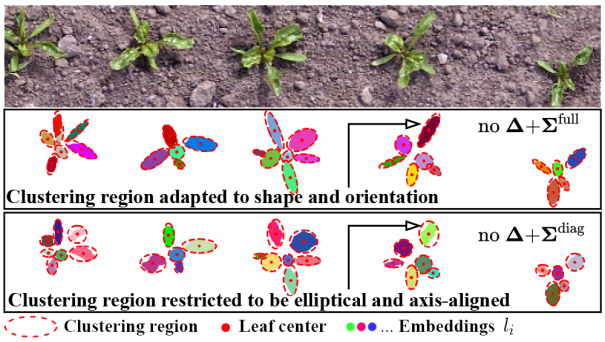

Figure 6: Image representation of the input image (top) after post-processing based on the network's predictions for crop leaves, when trained without offsets but $\boldsymbol{\Sigma}^{\mathrm{full}}$ (middle) or $\boldsymbol{\Sigma}^{\mathrm{diag}}$ (bottom).

These results convey that our predictions of full covariance matrices increase the segmentation performance and is superior to previous, more restricted representations [22].

5. Conclusion

In this work, we presented a novel vision-based approach to perform a simultaneous instance segmentation of crop leaves and plants using UAV-recorded images of real agricultural fields. Our proposed method generates two different image representations suitable to cluster individual crop leaves and plants within a predicted clustering region. We exploit these predictions to compute relevant basic phenotypic traits for individual crops in the field. Our thorough experimental evaluation using data from real agricultural fields suggests that our method outperforms multiple state-of-the-art approaches. We also show that our novel method to specify the clustering region based on full covariance matrices improves the overall performance in comparison with representations presented in related work.

Acknowledgments 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.

References

[1] Maxim Berman, Amal R. Triki, and Matthew B. Blaschko. The Lovász-Softmax Loss: A Tractable Surrogate for the Optimization of the Intersection-Over-Union Measure in Neural Networks. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[2] Bert De Brabandere, Davy Neven, and Luc Van Gool. Semantic Instance Segmentation with a Discriminative Loss Function. In *Deep Learning for Robotic Vision workshop, IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[3] Julien Champ, Adan Mora-Fallas, Hervé Goëau, Erick Mata-Montero, Pierre Bonnet, and Alexis Joly. Instance Segmentation for the Fine Detection of Crop and Weed Plants by Precision Agricultural Robots. *Applications in Plant Sciences*, 8(7):e11373, 2020.

[4] Dijun Chen, Ming Chen, Thomas Altmann, and Christian Klukas. Bridging Genomics and Phenomics. In *Approaches in integrative bioinformatics*, pages 299–333. 2014.

[5] Mark Everingham, Luc Van Gool, Christopher K.I. Williams., John Winn, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *Intl. Journal of Computer Vision (IJCV)*, 88(2):303–338, 2010.

[6] Wolfgang Förstner and Bernhard Wrobel. *Photogrammetric Computer Vision – Statistics, Geometry, Orientation and Reconstruction*. Springer Verlag, 2016.

[7] Mark S. Gockenbach. *Finite-Dimensional Linear Algebra*. CRC Press, 2011.

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proc. of the IEEE Intl. Conf. on Computer Vision (ICCV)*, 2017.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[10] Yotam Itzhaky, Guy Farjon, Faina Khoroshevsky, Alon Shpigler, and Aharon Bar-Hillel. Leaf Counting: Multiple Scale Regression and Detection Using Deep CNNs. In *Proc. of British Machine Vision Conference (BMVC)*, 2018.

[11] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv preprint*, abs/1412.6980, 2014.

[12] Victor Kulikov and Victor Lempitsky. Instance Segmentation of Biological Images using Harmonic Embeddings. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[13] Peter D. Lancashire, Hermann Bleiholder, T. van den Boom, P. Langelüddeke, Reinhold Stauss, Elfriede Weber, and A. Witzenberger. A Uniform Decimal Code for Growth Stages of Crops and Weeds. *Annals of Applied Biology*, 119(3):561–601, 1991.

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and Lawrence C. Zitnick. Microsoft COCO: Common Objects in Context. In *Proc. of the Europ. Conf. on Computer Vision (ECCV)*, 2014.

[15] Philipp Lottes, Jens Behley, Andres Milioto, and Cyrill 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.

[16] Federico Magistri, Nived Chebrolu, and Cyrill Stachniss. Segmentation-Based 4D Registration of Plants Point Clouds for Phenotyping. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2020.

[17] Chris McCool, Tristan Perez, and Ben Upcroft. Mixtures of Lightweight Deep Convolutional Neural Networks: Applied to Agricultural Robotics. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2017.

[18] Andres Milioto, Philipp Lottes, and Cyrill 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.

[19] Andres Milioto, Philipp Lottes, and Cyrill Stachniss. Real-time Semantic Segmentation of Crop and Weed for Precision Agriculture Robots Leveraging Background Knowledge in CNNs. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2018.

[20] Massimo Minervini, Hanno Scharr, and Sotirios A. Tsaftaris. Image Analysis: The New Bottleneck in Plant Phenotyping. *IEEE Signal Processing Magazine*, 32(4):126–131, 2015.

[21] Daniel Morris. A Pyramid CNN for Dense-Leaves Segmentation. In *Proc. of the Conf. on Computer and Robot Vision (CRV)*, 2018.

[22] Davy Neven, Bert De Brabandere, Marc Proesmans, and Luc Van Gool. Instance Segmentation by Jointly Optimizing Spatial Embeddings and Clustering Bandwidth. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[23] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. Personlab: Person Pose Estimation and Instance Segmentation with a Bottom-up, Part-based, Geometric Embedding Model. In *Proc. of the Europ. Conf. on Computer Vision (ECCV)*, 2018.

[24] Jean-Michel Papeand and Christian Klukas. 3-D Histogram-Based Segmentation and Leaf Detection for Rosette Plants. In *Proc. of the Europ. Conf. on Computer Vision (ECCV)*, pages 61–74, 2014.

[25] Mengye Ren and Richard S. Zemel. End-to-End Instance Segmentation and Counting with Recurrent Attention. *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[26] Eduardo Romera, José M. Alvarez, Luis M. Bergasa, and Roberto Arroyo. ERFNet: Efficient Residual Factorized ConvNet for Real-Time Semantic Segmentation. *IEEE Trans. on Intelligent Transportation Systems (ITS)*, 19(1):263–272, 2018.

[27] Bernardino Romera-Paredes and Philip Hilaire Sean Torr. Recurrent Instance Segmentation. *Proc. of the Europ. Conf. on Computer Vision (ECCV)*, abs/1511.08250, 2016.

[28] Hanno Scharr, Massimo Minervini, Andreas Fischbach, and Sotiriois A. Tsaftaris. Annotated Image Datasets of Rosette Plants. In *European Conference on Computer Vision. Zürich, Suisse*, pages 6–12, 2014.

[29] Hanno Scharr, Massimo Minervini, Andrew P. French, Christian Klukas, David M. Kramer, Xiaoming Liu, Imanol Luengo, Jean-Micel Pape, Gerrit Polder, and Danijela Vukadinovic. Leaf Segmentation in Plant Phenotyping: a Collation Study. *Machine Vision and Applications*, 27(4):585–606, 2016.

[30] Weinan Shi, Rick van de Zedde, Huanyu Jiang, and Gert Kootstra. Plant-part Segmentation Using Deep Learning and Multi-view Vision. *Biosystems Engineering*, 187:81–95, 2019.

[31] Gilbert Strang. *Linear Algebra and Learning from Data*. Wellesley-Cambridge Press, 2019.

[32] Jan Weyler, Andres Milioto, Tillmann Falck, Jens Behley, and Cyrill Stachniss. Joint Plant Instance Detection and Leaf Count Estimation for In-Field Plant Phenotyping. *IEEE Robotics and Automation Letters (RA-L)*, 2021.

[33] Yuli Wu, Long Chen, and Dorit Merhof. Improving Pixel Embedding Learning through Intermediate Distance Regression Supervision for Instance Segmentation. In *Proc. of the Europ. Conf. on Computer Vision (ECCV)*, pages 213–227, 2020.

[34] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2, 2019.