Supplement for 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

1. Detailed Layer Disposal

In Tab. 4 we provide a detailed overview of our network architecture which is based on ERFNet [26]. Let *chans* denote the number of feature maps per layer and *size* denote the width and height of each feature map.

The network follows the design of an encoder-decoder architecture that employs residual connections and factorized convolutions to remain efficient. In addition, it retains high accuracy on common benchmarks and outperforms a variety of competing network architectures. We emphasize that this network provides dense predictions. Thus, we select it as our backbone and adapt it to our task.

First, we feed the RGB input image into the encoder and forward its output to two different decoders, which are labeled as (a) and (b) in Fig. 2 and Tab. 4. The first decoder (a) predicts eight feature maps in its final output, see Tab. 4. These include the offset maps $\mathbf{\Delta}_L$ and $\mathbf{\Delta}_P$ that enforce pixels of individual crop leaves and plants to point into an instance-specific region around the center of the instance they belong to, see Sec. 3.2. In addition, this decoder predicts the feature maps $\mathbf{\Theta}_L$, $\mathbf{\Lambda}_L^1$, and $\mathbf{\Lambda}_L^2$ that we exploit to compute the inverse covariance matrix $\mathbf{\Sigma}_{L_k}^{-1}$ of each crop leaf, see Sec. 3.3. We employ the remaining three feature maps $\mathbf{\Theta}_P$, $\mathbf{\Lambda}_P^1$, and $\mathbf{\Lambda}_P^2$ to compute the inverse covariance matrix $\mathbf{\Sigma}_{P_j}^{-1}$ of each crop plant. The second decoder (b) predicts two feature maps in its final output. We exploit the predicted feature map $\mathbf{S}_L$ to recover the instance centers of crop leaves and $\mathbf{S}_P$ to recover the instance centers of crop plants, as described in Sec. 3.4.

Table 4: Detailed layer disposal of our proposed network.

	Layer	Type	chans	size
	0	RGB Input Image	3	1024×512
Encoder	1	Downsampler Block [26]	16	512×256
	2	Downsampler Block	64	256×128
	3-7	5×Non-bt-1D [26]	64	256×128
	8	Downsampler Block	128	128×64
	9	Non-bt-1D (dilated 2)	128	128×64
	10	Non-bt-1D (dilated 4)	128	128×64
	11	Non-bt-1D (dilated 8)	128	128×64
	12	Non-bt-1D (dilated 16)	128	128×64
	13	Non-bt-1D (dilated 2)	128	128×64
	14	Non-bt-1D (dilated 4)	128	128×64
	15	Non-bt-1D (dilated 8)	128	128×64
	16	Non-bt-1D (dilated 16)	128	128×64
(a) Decoder	17 (a)	Deconvolution	64	256×128
	18-19 (a)	2×Non-bt-1D	64	256×128
	20 (a)	Deconvolution	16	512×256
	21-22 (a)	2×Non-bt-1D	16	512×256
	23 (a)	Deconvolution	8	1024×512
(b) Decoder	17 (b)	Deconvolution	64	256×128
	18-19 (b)	2×Non-bt-1D	64	256×128
	20 (b)	Deconvolution	16	512×256
	21-22 (b)	2×Non-bt-1D	16	512×256
	23 (b)	Deconvolution	2	1024×512

2. Comparison of Score Maps

At inference, we recover the centers of crop leaves and plants based on the predicted feature maps $\mathbf{S}_L$ and $\mathbf{S}_P$, respectively. The map $\mathbf{S}_L$ contains high scores at pixel locations $\mathbf{x}_i$ where the associated spatial embedding $\mathbf{l}_i = \mathbf{x}_i + \Delta\mathbf{l}_i$ has a high confidence score under the Gaussian in Eq. (1). This occurs if the embedding $\mathbf{l}_i$ is close to the center of a crop leaf. Accordingly, the map $\mathbf{S}_P$ has high scores at pixel locations where the corresponding embedding $\mathbf{p}_i = \mathbf{l}_i + \Delta\mathbf{p}_i$ has a high confidence score under the Gaussian in Eq. (2), *i.e.*, it is close to the center of a crop plant. Thus, the appearance of these maps depends on our network's prediction for the offset maps $\mathbf{\Delta}_L$ and $\mathbf{\Delta}_P$. In Fig. 7 we show the predicted maps $\mathbf{S}_L$ and $\mathbf{S}_P$ in case we train the network with or without the offsets $\mathbf{\Delta}_L$ and $\mathbf{\Delta}_P$.

In the first case, almost all spatial embeddings $\mathbf{l}_i$ are close to the center of their associated crop leaf. Thus, the map $\mathbf{S}_L$ contains a high score at each pixel location $\mathbf{x}_i$ which belongs to a crop leaf. The same holds for the embeddings $\mathbf{p}_i$ and the predicted map $\mathbf{S}_P$.

In the second case, all spatial embeddings are equal to their pixel location $\mathbf{x}_i$ since we explicitly set

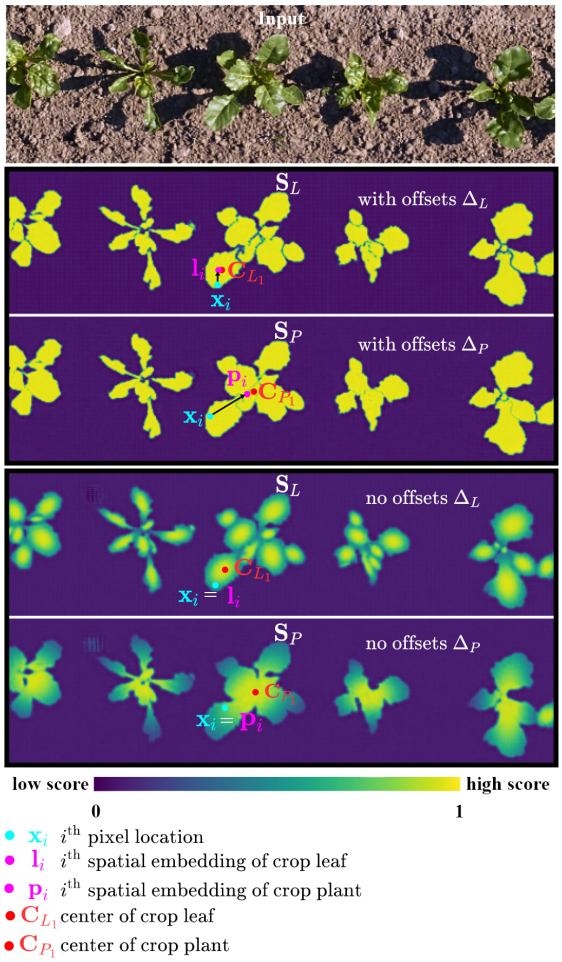

Figure 7: Comparison between the predicted maps $\mathbf{S}_L$ and $\mathbf{S}_P$ of an input image (top) for different optimization procedures of our network. In the first case (top), the network translates pixels that belong to an instance towards their associated center and predicts small clustering regions. In the second case (bottom), the network predicts no offsets but adapts the clustering region to the shape and orientation of an instance.

$\Delta\mathbf{l}_i = \Delta\mathbf{p}_i = 0$. Thus, the map $\mathbf{S}_L$ contains high scores only at pixel locations that are nearby the actual center of a crop leaf. In contrast, we observe lower scores for all pixels that belong to a crop leaf but are located at a certain distance from the corresponding crop leaf center. The same effect holds for the map $\mathbf{S}_P$, which has high scores at pixel locations that are close to the center of a crop plant.

However, we state that these cases reflect different options of our network to optimize the objectives in Eq. (3) and Eq. (4) that result in different maps $\mathbf{S}_L$ and $\mathbf{S}_P$.

First, the network can translate pixels towards their desired instance centers and predict small clustering regions around an object's center to achieve high scores under the Gaussian in Eq. (1) and Eq. (2). This results in maps $\mathbf{S}_L$ and $\mathbf{S}_P$ that have a similar visual appearance, as shown in the center of Fig. 7.

Second, the network can also predict minor offsets but adapt the clustering regions represented by covariance matrices to the shape and orientation of an object. With regard to the objectives in Eq. (3) and Eq. (4) it is sufficient to predict covariance matrices that achieve confidence scores > 0.5 under the Gaussians in Eq. (1) and Eq. (2) for all embeddings which belong to an instance. This results in maps $\mathbf{S}_L$ and $\mathbf{S}_P$ where the scores decrease with increasing distance to the instance center, as shown at the bottom of Fig. 7. Here, both maps appear very differently.

Both previously described options are valid w.r.t. to the objectives in Eq. (3) and Eq. (4). This emphasizes that we need specific maps to recover the centers of crop leaves and plants, respectively. Thus, a single map is not sufficient to recover instance centers. In both cases, the map $\mathbf{S}_L$ is appropriate to recover instance centers of crop leaves, and $\mathbf{S}_P$ is well-suited to recover instance centers of crop plants.

3. Qualitative Results CVPPP LSC

We provide more qualitative results of our approach on the popular CVPPP LSC in Fig. 8. We follow best practice and use sequence A1 with the highest number of published results. This sequence contains 128 labeled images for training and 33 test images with a size of $530\,\text{px} \times 500\,\text{px}$ each. The task of this competition is to segment each leaf of a single plant recorded in a laboratory environment. In contrast, we explicitly designed our network to perform a simultaneous instance segmentation of crop leaves and plants on images of real agricultural fields that contain an arbitrary number of plants. Thus, the LSC addresses a less complex problem. In Fig. 8 we show that our approach achieves an accurate leaf segmentation performance that outperforms competing methods, see Tab. 3.

To the best of our knowledge, no publicly available competition covers a simultaneous instance segmentation of crop leaves and plants on agricultural fields as targeted by our approach. Thus, we report the performance of our method on the CVPPP LSC in order to present a quantitative analysis with published result.

4. Baselines and Additional Qualitative Results

In Fig. 8 we provide more qualitative results of our method on our dataset in comparison with two baselines. We emphasize that our approach performs a simultaneous instance segmentation of crop leaves and plants in a single network, which associates each detected leaf with a specific crop plant. This is a challenging task, since the number of crops in the field as well as the number of leaves per plant is highly variable. To the best of our knowledge, there is no method which explicitly models a simultaneous instance segmentation of individual crop leaves and plants on real agricultural fields. The method proposed by

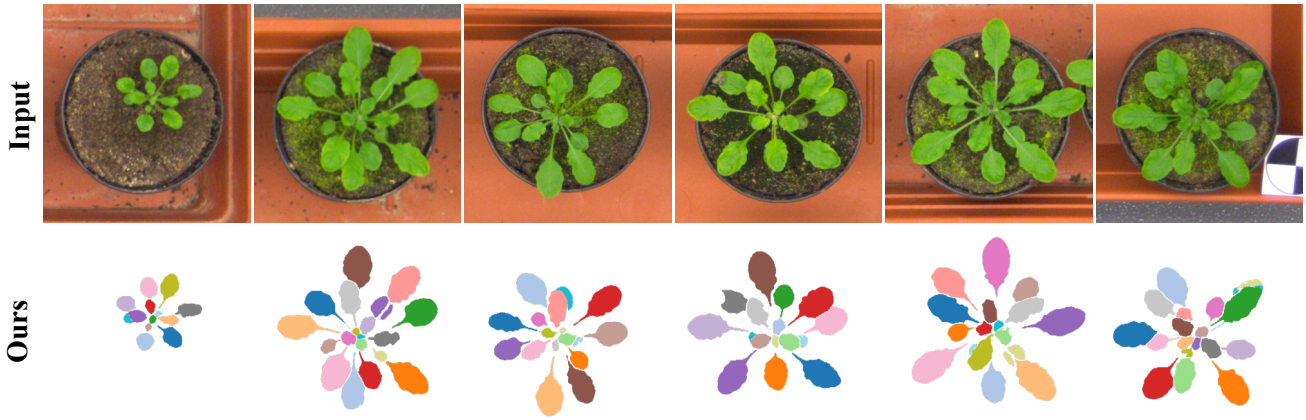

Figure 8: Qualitative results of our approach on the CVPPP LSC. We show the input images (top row) and the corresponding leaf instance segmentation of our method (bottom row). We represent each leaf in a color-encoded fashion.

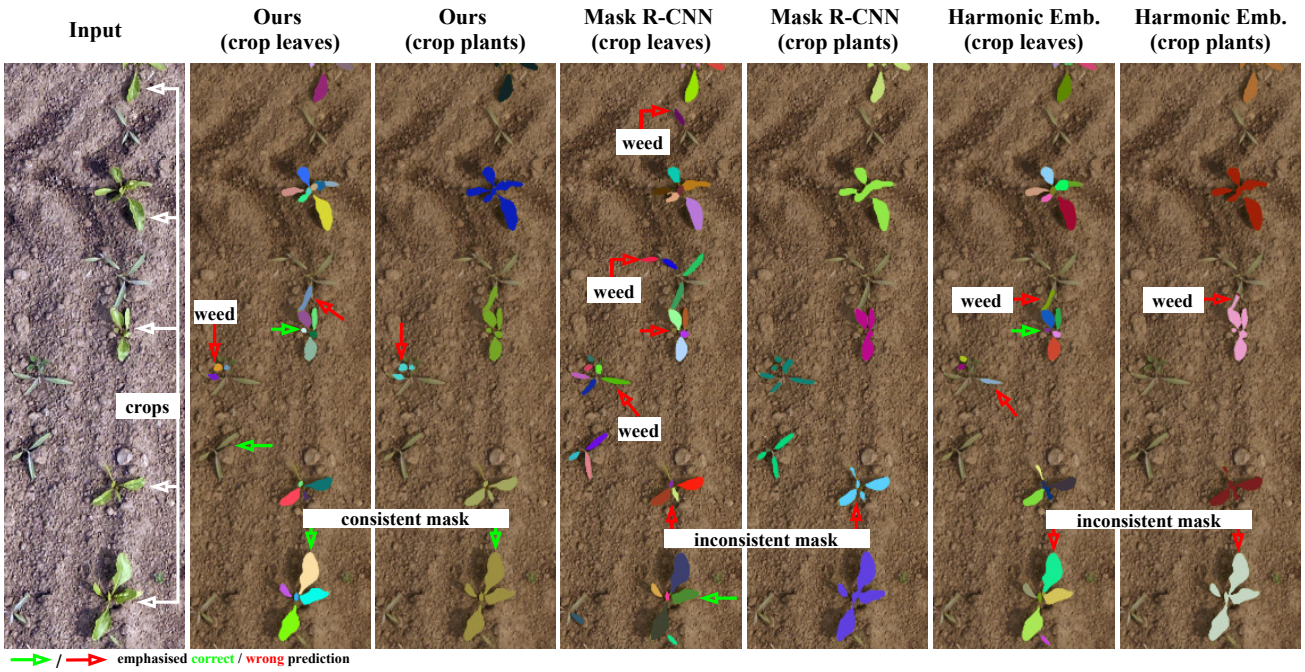

Figure 9: Qualitative results of our approach on our dataset in comparison with two baselines. Note that we show a cropped and enhanced input image for better visibility of crops and weeds.

Weyler *et al*. [32] targets a similiar objective but provides only coarse bounding boxes for each plant and coarse keypoint locations for each leaf. Thus, we compare our method to different approaches that achieve state-of-the-art results for the task of instance segmentation, *i.e*., Mask R-CNN [8] and Harmonic Embeddings [12]. The former approach achieves high performance on a variety of benchmarks and is still one of the most used methods for instance segmentation. In contrast, the latter method is explicitly designed to achieve high accuracy on biological images. However, both methods do not allow to perform a simultaneous instance segmentation of crop leaves and plants in a single network. Thus we train two networks for each of the proposed baselines. We train the first network only on instances of crop leaves and the second network only on instances of crop plants. Thus, each of the two networks is an expert for the task of crop leaf instance segmentation or crop plant instance segmentation. At inference, we apply each network independently to the input data and report the results.