PhenoBench: A Large Dataset and Benchmarks for Semantic Image Interpretation in the Agricultural Domain

Jan Weyler, Federico Magistri, Elias Marks, Yue Linn Chong, Matteo Sodano, Gianmarco Roggiolani, Nived Chebrolu, Cyrill Stachniss, and Jens Behley

Abstract—The production of food, feed, fiber, and fuel is a key task of agriculture, which has to cope with many challenges in the upcoming decades, e.g., a higher demand, climate change, lack of workers, and the availability of arable land. Vision systems can support making better and more sustainable field management decisions, but also support the breeding of new crop varieties by allowing temporally dense and reproducible measurements. Recently, agricultural robotics got an increasing interest in the vision and robotics communities since it is a promising avenue for coping with the aforementioned lack of workers and enabling more sustainable production. While large datasets and benchmarks in other domains are readily available and enable significant progress, agricultural datasets and benchmarks are comparably rare. We present an annotated dataset and benchmarks for the semantic interpretation of real agricultural fields. Our dataset recorded with a UAV provides high-quality, pixel-wise annotations of crops and weeds, but also crop leaf instances at the same time. Furthermore, we provide benchmarks for various tasks on a hidden test set comprised of different fields: known fields covered by the training data and a completely unseen field. Our dataset, benchmarks, and code are available at https://www.phenobench.org.

1 Introduction

The agricultural production of food, feed, fiber, and fuel has to cope with several challenges in the upcoming decades. The world population is increasing, yet the availability of arable land is limited or even decreasing, climate change increased uncertainties in crop yield, and we observe substantial losses in biodiversity [18]. At the same time, agricultural practices need to be more sustainable and have to reduce the use of agrochemical inputs, *i.e.*, herbicides and fertilizers that potentially negatively impact yield [32] and the environment.

Robots and drones using vision-based perception systems could help with these challenges by offering tools to make better, more sustainable field management decisions and providing supporting tools for breeding new varieties of crops by estimating plant traits in a reproducible manner [72]. Such visual perception systems enable the development of agricultural robots that can support the monitoring of fields and replace labor-intensive tasks such as manual weeding [89]. Additionally, they potentially enable more targeted crop management, where agrochemicals are applied precisely and only where needed, thereby reducing the negative effects on the environment [53], [82].

With the advent of deep learning for visual perception [41], [49], the field of computer vision has made tremendous progress in image interpretation, achieving remarkable results in several domains. Datasets and associated benchmarks [14], [52], [66] were essential for achieving this progress as they provide a testbed for developing novel algorithms but also provided the necessary data to tackle novel tasks. Progress can be tracked quantitatively with metrics that measure the performance of developed approaches against benchmarks using hidden test sets. Novel tasks with increasing complexity drive the progress of the field by posing novel challenges for the community.

In this paper, we aim to provide a large dataset together with benchmarks for semantic interpretation under real-field conditions enabling similar progress in the agricultural domain. We target multiple tasks: semantic segmentation, panoptic segmentation, plant detection, leaf detection, and the novel task of hierarchical panoptic segmentation that provides a coarse-to-fine interpretation of plants.

For this purpose, we recorded high-resolution images with unmanned aerial vehicles (UAV) of sugar beet fields under natural lighting conditions over multiple days, capturing a large range of growth stages. We annotated these images with dense, pixel-wise annotations to identify sugar beet crops and weeds at an instance level, as needed for semantic segmentation and plant-level instance segmentation tasks. Additionally, we labeled leaf instances of crops to enable the investigation of leaf instance segmentation (see Fig. 1). Furthermore, we provide temporal association of plant instances over the different dates, which allows to identify individual plants at different growth stages.

The combination of plant-level and leaf-level annotations enable the investigation of novel tasks needed for a holistic semantic interpretation in the agricultural domain. One such task is the hierarchical panoptic segmentation that

- *J. Weyler, F. Magistri, E. Marks, Y.L. Chong, M. Sodano, G. Roggiolani, and J. Behley are with the Center for Robotics, University of Bonn, Germany. E-mails: {firstname.lastname}@igg.uni-bonn.de*
- *N. Chebrolu is with the University of Oxford, UK. E-mail: nived@robots.ox.ac.uk*
- *C. Stachniss is with the Center for Robotics, University of Bonn, Germany, the University of Oxford, UK, and the Lamarr Institute for Machine Learning and Artificial Intelligence, Germany. E-Mail: cyrill.stachniss@igg.uni-bonn.de*

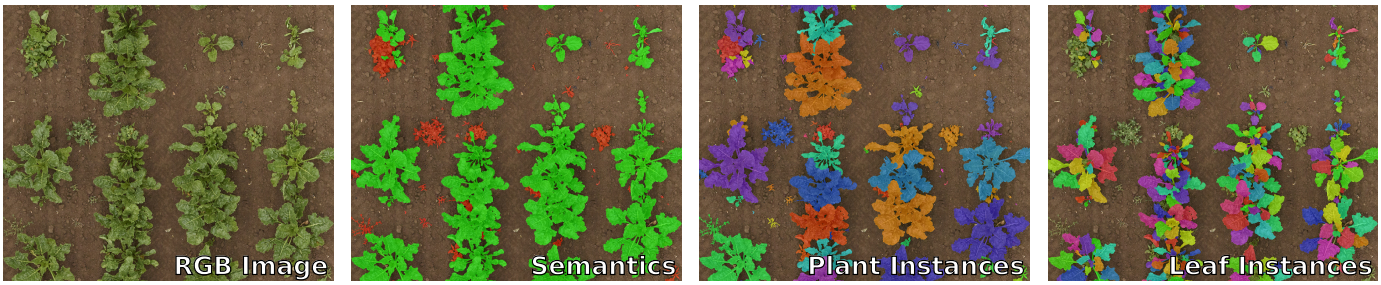


Fig. 1. Our dataset, called *PhenoBench*, provides dense semantic plant-level instance annotations (shown by different colors) of sugar beet crops and weeds (green and red in the semantics) and leaf-level instance annotations of crops (different colors correspond to different instances) for high-resolution images recorded with a UAV. The dataset consists of images collected at different times during a growing season, which captures various growth stages of plants.

targets to segment individual leaves and assign them to their associated plant instance to predict the total number of leaves per plant. Plant scientists and breeders commonly assess this information to describe the growth stage of individual plants, which is also linked to yield potential and plant performance [45]. However, this in-field assessment is nowadays done manually outside greenhouses, which is laborious and time-consuming [62]. Thus, developing vision systems to assess these properties per plant automatically is essential for large-scale, sustainable crop production.

Our provided data shows distinct challenges in terms of plant variation and overlap between different plant and leaf instances that are distinct in the agricultural domain. Such challenges are seldomly addressed by general segmentation approaches prevalent in man-made environments, as shown by our experimental results, where we challenged several state-of-the-art approaches but also provide results for more domain-specific approaches for the agricultural domain.

In summary, our main contributions are:

- We present a large dataset for plant segmentation providing accurate instance annotations at the level of plants and leaves.
- We provide benchmark tasks on a hidden test set for evaluating semantic, instance, and panoptic segmentation, and detection approaches targeted at plants enabling reproducible and unbiased evaluation of novel plant perception approaches.
- We provide baseline results for general and domain-specific models for plant and leaf detection, but also semantic, instance, and panoptic segmentation.

We believe that the effort in generating high-quality annotations and establishing reliable benchmarks for multiple tasks with a hidden test set will accelerate progress in semantic perception of agricultural fields and potentially lead to novel avenues of research in this important domain. We make our dataset and benchmarks[1], code for visualizing predictions and computing metrics[2], and baselines[3] with code, checkpoints, and predictions publicly available.

2 Related Work

In recent years, dense, pixel-wise semantic interpretation of images, *i.e.*, semantic, instance, and panoptic segmentation [38], made rapid progress due to advances in deep learning [49], but also thanks to the availability of large-scale datasets for object detection [20], [21], [52], semantic segmentation [14], [66], instance segmentation [52], and lately panoptic segmentation [14], [52], [66].

Despite the availability of large datasets in man-made environments, the agricultural domain faces different challenges, such as large intra-class variability due to plant growth. Thus, there has been interest in large datasets to enable studying perception in the agricultural domain [57]. However, accurately dense annotated and large agricultural datasets in combination with reproducible benchmarks on a hidden test set are still missing today, see Tab. 1.

In particular, the crop/weed field image dataset (CWFID) by Haug *et al.* [30] is one of the first semantic segmentation datasets that provides pixel-level annotations of semantics for plants, *i.e.*, sugar beets and weeds using a multispectral camera. Lameski *et al.* [44] also provides a dataset for crops, *i.e.*, carrots and weed segmentation. CVPPP [1], [61] is one of the first datasets providing annotations for leaves in images of individual tobacco and arabidopsis plants recorded in a lab environment, which is also the basis for a series of workshops and competitions hosted at CVPR and ICCV. The dataset by Chebrolu *et al.* [7] provides images of sugar beets and weeds recorded by a ground robot under real field conditions with a ground sampling distance (GSD) of $0.3\,\frac{\text{mm}}{\text{px}}$ and provides annotations for semantic segmentation. Similar to our dataset, the WeedMap dataset [79] provides imagery of UAVs covering a large field with sugar beets and weeds. In contrast to our dataset, where we provide the original camera data, WeedMap first generated orthophotos via bundle adjustment. While we considered this option, we noticed that the lack of a detailed elevation model usually leads to artifacts on the boundaries of the plants. Additionally, the images of WeedMap have a coarse GSD between $8.2\,\frac{\text{cm}}{\text{px}}$ and $13\,\frac{\text{cm}}{\text{px}}$ while our images have a GSD of $1\,\frac{\text{mm}}{\text{px}}$ to assess detailed information for individual plants. The Sunflowers dataset [22] provides images collected with a multi-spectral sensor providing RGB and near-infrared images from a ground robot. The Agriculture-Vision dataset [13] contains aerial images with a coarse GSD between $10\,\frac{\text{cm}}{\text{px}}$ and $20\,\frac{\text{cm}}{\text{px}}$ with corresponding annotations that covers rather large areas but not individual plants, *e.g.*, regions with nutrient deficiencies and weed clusters.

1. https://www.phenobench.org
2. https://github.com/PRBonn/phenobench
3. https://github.com/PRBonn/phenobench-baselines

Dataset	#Images	Image Size	Crop			Weed		Field?	Hidden Test Set?
			Sem.	Inst.	Leaves	Sem.	Inst.		
CWFID [30]	60	1291 × 966	✓			✓		✓	
CVPPP [1], [61]	1,311	2048 × 2448[1]			✓				✓
Carrot-Weed [44]	39	3264 × 2448	✓			✓		✓	
Sugar beets [7]	280	1296 × 966	✓			✓		✓	
WeedMap [79]	1,670	480 × 360	✓			✓		✓	
Carrots-Onion [3]	40	2464 × 2056	✓			✓		✓	
Oil Radish [65]	129	1600 × 1600	✓			✓		✓	
Sunflower [22]	500	1296 × 966	✓			✓		✓	
GrowliFlowers [36]	2,198	448 × 368	✓	✓	✓			✓	
CropAndWeed [81]	8,034	1920 × 1088	✓			✓		✓	
PhenoBench (Ours)	2,872	1024 × 1024	✓	✓	✓	✓	✓	✓	✓

TABLE 1. Comparison of datasets in the agricultural domain providing *dense pixel-wise annotations*. For the crop and weed, we indicate if semantic segmentation (Sem.), plant instances (Inst.), and leaf instances (Leaves) are densely annotated. We also record if the dataset was recorded under field conditions, as opposed to under lab conditions (Field?). Furthermore, we note if there is a hidden test set, such that approaches do not have access to test set labels (Hidden Test Set?). [1]We report maximum image size, as it ranges from $441\,\text{px} \times 441\,\text{px}$ to $2048\,\text{px} \times 2448\,\text{px}$.

More recently, the GrowliFlowers dataset [36] provides images recorded with a UAV showing multiple growth stages of cauliflowers. While we recorded images on three dates roughly a week apart, this dataset contains images captured on four different dates, also roughly a week apart. Therefore, it captures an extended period of one month.

Lately, the CropAndWeed dataset [81] provides RGB images taken close to the field canopy showing a large variety of crops and weeds. While the number of annotated images is large, the pixel-wise annotations have been semi-automatically annotated exploiting a pre-segmentation via a deep neural network to lower the annotation effort. However, this sometimes leads to incomplete annotations and notable annotation artifacts. Also in our experience, we noted that correcting annotations is quite tedious and can counter-intuitively lead to even larger annotation effort as boundary regions generated using contemporary segmentation approaches almost always need to be corrected, which is also the part that takes most of the annotation time.

The recently published RumexLeaves dataset [27] provides fine-grained annotations of leaves of the Rumex obtusifolius L., which is a problematic weed in grasslands. Besides the leaf annotations of this particular plant, the dataset also provides more fine-grained vein and stem annotations that allows to get insights into the plant physiology corresponding to traits relevant for plant phenotyping.

Besides the aforementioned closely related datasets that also provide dense pixel-wise annotations, there have been recently also several datasets in the agricultural domain released for wheat detection [16], localization and mapping [34], [69], image classification of weed species [67], detection for phenotyping [58], crop row detection [96], or fruit detection [78]. Additionally, there are a small number of available datasets for semantic interpretation of 3D agricultural data [19], [35], [80].

While recent interactive labeling approaches, like SegmentAnything [39], can certainly speed-up labeling of instance masks with only weak annotations delivering compelling results, we target to generate a reliable and high-quality dataset and corresponding benchmark. Therefore, we resort to manual annotations from scratch, which entailed a rigorous correction and verification procedure to ensure accurate and consistent segmentation masks.

In contrast to the aforementioned datasets, which are great starting points for research, our dataset shows an unique level of annotations, including semantic and instance masks for crops and weeds of an overall larger number individual plants (see Tab. 1). Furthermore, we provide temporally consistent instance ids of crops that allow to identify individual plants over multiple dates. Note that our dataset provides large images with multiple completely visible plants, which is not always the case for other pixel-wise annotated datasets [36], [61]. Lastly, we enable comparable and reproducible results with the provided benchmarks on a hidden test set, *i.e.*, labels are not released and the predictions are evaluated on a server via CodaLab [68].

3 Our Dataset

In this section, we present our setup for data collection, explain the labeling process, and provide statistics to show the variability of the data.

3.1 Data Collection

Our dataset provides RGB images in real field conditions recorded by an UAV equipped with a high-resolution camera that captures imagery of the field. For recording the data, we employed a DJI M600 and used the PhaseOne iXM-100 camera with a $80\,\text{mm}$ RSM prime lens mounted on a gimbal to obtain motion-stabilized RGB images at a resolution of $11\,664\,\text{px} \times 8750\,\text{px}$. The UAV was flying at a height of approx. $21\,\text{m}$, resulting in a GSD of $1\,\frac{\text{mm}}{\text{px}}$. For covering the entire field, we use the DJI Ground Station Pro app to plan a flight that covers the field row-wise. We set the forward overlap between consecutive images by motion vector at $75\,\%$ and the side overlap between images placed in neighboring rows at $50\,\%$. Each image is geo-referenced by using the on-board GNSS.

We performed three missions roughly a week apart to capture different growth stages of the plants. More specifically, we performed the flights on May 15, May 26, and June 6 in 2020. Additionally, we used the same sensor setup to record images at four different points in time in 2021 on a

Fig. 2. Variability in overlap and illumination of plants at the same part of the field on different recording dates. Theses examples show the variation in growth stages ranging from 4 leaf stage (early growth stage) to plants with over 20 leaves (later growth stage) and the variety of illuminations with sunny (left) and overcast (right) weather conditions.

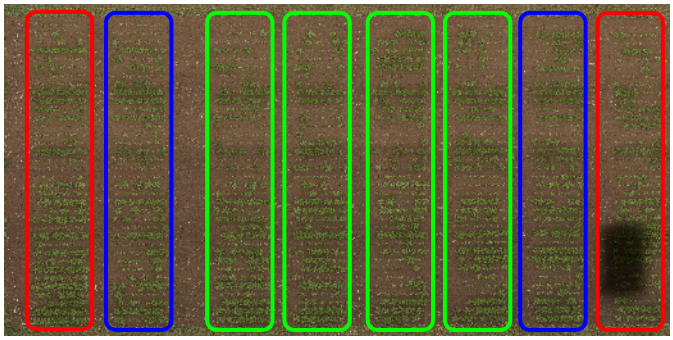


Fig. 3. Orthophoto of the field recorded in 2020 and our spatial separation into rows for training (green), validation (blue), and testing (red). Due to the geo-referencing of the images, we extracted the same rows on each of the dates.

different field: May 20, May 28, June 1, and June 10. As the data was captured in the open field, we have a variety of different lighting conditions with sunny and also overcast weather, as shown in Fig. 2, which significantly changes the visual appearance of the plants.

From the approximately $1300\,\mathrm{m}^2$ sugar beets field located at the Campus Klein-Altendorf farm between Meckenheim and Rheinbach, Germany (50°37′.51N, 6°59′.32E), we selected eight crop rows that were covered by the recording mission. To have a clear spatial separation between the train and test set, we selected four crop rows for extracting training images, two crop rows for validation, and two crop rows for testing purposes as shown in Fig. 3. Additional data recorded in 2021 is only included in the test set to evaluate also the performance in a setting of an unseen field with the same crop but potentially different weeds.

Specifically, the sugar beet field contains a mixture of two different crop varieties, *i.e.*, BTS 440 and Celesta KWS that are both from distinct agro-seed companies and differ in their properties regarding a beet's mass and sugar yield. Furthermore, we observe six weed varieties that are most prominent in the field, *i.e.*, Chenopodium album, Polygonum aviculare, Thlaspi arvense, Persicaria lapathifolia, Bilderdykia convolvulus, and Polygonum hydropiper.

The field belongs to a farm of the University of Bonn located at the Campus Klein-Altendorf. This allows us to conduct field studies and to study perception systems under varying conditions with respect to application of herbicides, which leads to different scenarios with fully (conventional), partial (80% herbicides), and non-herbicided field conditions, as shown in Fig. 4. In conventional farming and field management operations, such conditions with less or no herbicides are usually not observable. While keeping most of the other field parameters constant, this makes our field setup distinct to other larger datasets, such as GrowliFlowers [36] that recorded data only under conventional field management conditions with only a very few weeds.

Fig. 4. Varying conditions of the field recorded at different locations, which are treated with different amounts of herbicides. From left to right: Fully-herbicided, partially-herbicided, and non-herbicided field conditions recorded at the same day.

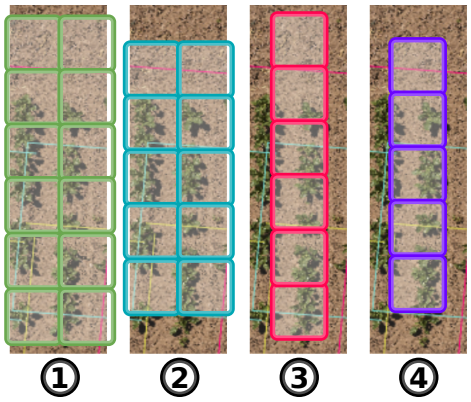


Fig. 5. Extracted tiles per iteration such that a row is densely covered with tiles to ensure that all plants are completely visible in at least one tile. Annotations of tiles are transferred between iterations and aggregated in the global image I_g.

3.2 Labeling Process

The full-sized images, which we denote as global images, I_g, are challenging to annotate due to their large size of $11\,664\,\mathrm{px} \times 8750\,\mathrm{px}$. To parallelize the labeling process and ensure no plant is missed, we extracted from I_g overlapping patches, I_p, of size $2000\,\mathrm{px} \times 2000\,\mathrm{px}$. We extracted multiple iterations of overlapping patches such that we always have in one of the resulting four tilings complete plants visible, *c.f.* Fig. 5. As we ensure that each plant is fully visible in at least one of the patches, we instructed our annotators to label only completely visible plants in I_p.

For labeling the plants and leaves at the same time, we developed a novel tool to enable a hierarchical annotation of the images. Please see the supplement for a more detailed description of the labeling tool and the provided features.

We first labeled the plant instances of sugar beet crops and weeds, which was completed by 9 annotators investing a total of 800 h. Each iteration was validated and corrected before we transferred the annotations to the global images I_g. Then, the next iteration is started with the transferred labels copied to the respective patches I_p, and these steps were repeated till the final fourth iteration.

Split	#imgs	#crops	#weeds	#leaves
Train	1,407	11,875	8,141	71,264
Validation	772	6,482	3,926	35,503
Test	693	6,201	4,291	33,935
Unlabeled	129,000	–	–	–

TABLE 2. Dataset statistics of the provided splits. Note that we have a hidden test set, *i.e.*, we have a server-sided evaluation [68]. We additional provide unlabeled data of the fields to enable studying of self-supervised pre-training.

Annotation of a single patch I_p ranged from approx. 1 h for earlier growth stages to 3.5 h for later growth stages where plants had significant overlap. In sum, we annotated 705 patches over all dates and crop rows.

After the plant instances were labeled, we had 5 annotators labeling leaf instances. Annotators were tasked with identifying crop leaves and annotation of a patch I_p took approx. 1 h to 2 h depending on the number of visible crops. With the masking of plant instances provided by our annotation tool, we ensure that we have consistent leaf labels that are inside the crop instance. Thus, it is possible to associate each leaf instance with its corresponding crop based on the plant instance annotations.

To ensure high-quality, accurate annotations of plants and leaves, we furthermore had an additional round of corrections performed by four additional annotators that revised the annotations. More details on our quality assurance process is provided in the supplementary material.

In total, we had 14 annotators who invested 1,400 h of annotation work and roughly 600 h invested into additional validation and refinement, leading to an overall labeling effort of approximately 2,000 h.

3.3 Temporal Alignment

As we recorded images in the same geographical location, we can furthermore provide temporally aligned plant instances, which enables the study of individual plant growth. By matching the occurrences of the same plants in different recordings we ensure that each crop plant has a unique instance id throughout our whole dataset.

To this end, we exploit the positions delivered by the RTK GNSS of the drone as initial guesses for a bundle adjustment procedure to determine the pose of the camera for each captured image in a global reference frame. This allows us to project the crop center locations, computed as the centroid of the plant pixels, of plants appearing in all images of a mission into a common plane.

As the estimated poses of the camera are not completely free of noise, we use Hungarian matching [42] based on the distances of crop centers to robustly associate instances of the same plant appearing in different images. To account for new crop instances but also missing crop instances, we only associate crop centers, when their distance is below a threshold of 15 cm, which was determined empirically. We experimented with using GNSS poses to associate crop instances between different missions collected at different points in time, but found the inaccuracies of the localization to be too high for our purpose. We, therefore, manually associated around 10 plants between the different missions and used these datapoints to compute a transformation between each mission using a least squares approach. Given those transformations we then associated the crop ids again by projecting them onto a common plane and matching them by the Hungarian algorithm. Finally, we validated the temporal alignment by visualizing the matches between missions at different points in time.

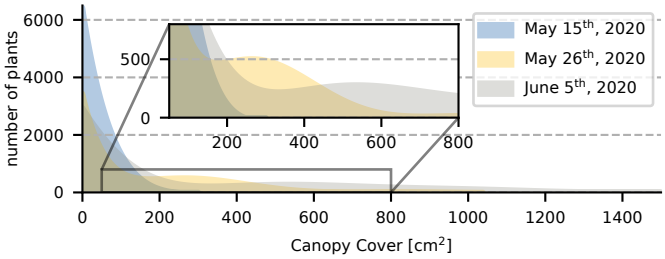


Fig. 6. Distribution of crop sizes in terms of canopy cover in cm^2 for mostly visible plants in the training and validation set.

3.4 Dataset Statistics

We finally extracted from the global images I_g smaller images of size $1024\,\text{px} \times 1024\,\text{px}$ to ensure that we have images containing complete crops at later growth stages, but also provide context such as the crop row structure. More specifically, we use the an overlap of 50% between extracted patches to ensure that plants in later growth stages are at least 50% visible in the extracted patches.

Tab. 2 shows an overview of the number of extracted images for the different splits from the earlier described train/validation/test rows, the number of crop instances, the number of crop leaves, and the number of weed instances annotated. Note that only the test data includes data from 2020 and 2021. As we ensured that we have completely annotated plants, we are able to generate a visibility map and differentiate between mostly visible plants with at least 50 % visible pixels and partially visible plants. Note that we provide a rather large validation set to allow researchers to conduct conclusive ablations studies.

In addition to the labeled data, we also provide unlabeled data from all fields, which can be exploited for pre-training, semi-supervised, or unsupervised domain adaptation, which we see as promising future avenue of research.

As motivated earlier, we recorded images under real-world conditions of real agricultural fields leading to a diverse range of plant appearances due to varying growth stages. The crops are affected by different soil conditions leading to a variety of growth stages even on images of the same date. This intra-class variability of the crops poses an interesting challenge for learning approaches that have to correctly segment or detect small but also large crops at the same time. The extra data from a different field captured in 2021 leads to even greater diversity of recording conditions, which is a common challenge in the agricultural domain.

Additionally, we observe a large variability in terms of overlap between plants. They are clearly separated at the beginning of the recording campaign but show a considerable overlap at the last recording date. Fig. 2 shows the same area of the field over the course of three weeks showing the variation in terms of growth stage but also the overlap between crops.

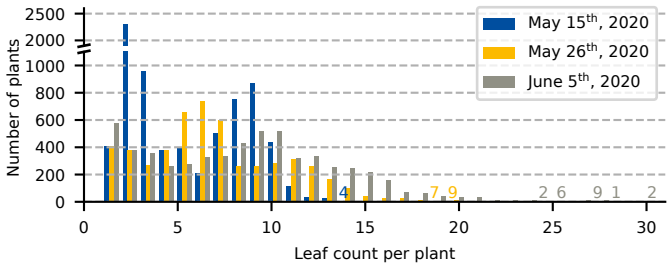


Fig. 7. Distribution of leaf count of mostly visible plants in the training and validation set.

In Fig. 6, we provide an overview of the plant sizes per data collection day in terms of the area covered by the plant instances that shows the diversity in terms of growth stages. While on May 20th plants with a small coverage are predominately present, the plant area of plants naturally increased in the following weeks. On May 26th, the amount of larger plants increases. At the latest date, June 5th, the amount of larger plants further increases and the distribution gets more long-tailed as now all plants directly compete for space, which is also visually visible from the larger overlap between neighboring plants. Thus, only few plants are able to develop a larger canopy cover.

Finally, we present in Fig. 7 the distribution of leaves per plant per data collection day of completely visible plants in the training and validation split. Similar to the trends for the canopy cover, we can also observe an increase in terms of the number of leaves over time. On May 20th, most of the plants are still in the two-leaves stage with only a few plants in the later development with more than 10 leaves. Note that some leaves are also so-called germ leaves that are later replaced by the real leaves. The peak in the leaf count shifts to the right on May 26th as the sugar beet plants develop more leaves in later growth stages. On the last data collection date, June 5th, the distribution of leaves gets more long-tailed as now larger plants are competing for space. At this stage, however, it's also more likely that leaves are covered by other leaves, since we observe the field from a UAV. Thus, the true number of leaves is not observable.

Overall, we annotated 583 unique crop plants at potentially different growth stages growing under real-world conditions in the open field. Thus, the individual plant growth is affected by the weather conditions and the soil quality that changes over the whole field. As noted before, the visual appearance changes between different plants but also can have substantial differences due to the natural plant growth. More specifically, 496 plants appear in all three dates, 15 plants in only two of the dates, and 72 plants only at a single point in time, which is caused by the conventional field management operations or natural growing conditions.

4 Benchmarks

In this section, we present the benchmark tasks that we provide together with the dataset. These tasks cover different aspects of a perception system for the crop production domain in agriculture. While we cover classical, well-established tasks, we also want to provide a novel task of hierarchical panoptic segmentation that provides a complete picture of the plant structure.

Approach	mIoU	IoU		
		Crop	Weed	Soil
ERFNet [76]	85.98	94.30	64.37	99.28
DeepLabV3+ [9]	85.97	94.07	64.59	99.25

TABLE 3. Baseline results for semantic segmentation on the test set.

We provide metrics on the test set of our dataset including data from *known* and *unknown* fields for all investigated baseline approaches. Note that we provide more details on the training setup, including hyperparameters and qualitative results, in the supplement. We furthermore will provide code for the baselines in our code release. In the supplement, we furthermore provide qualitative results together with more fine-grained quantitative results differentiating between the different fields of the test set.

4.1 Semantic Segmentation

Task description. Semantic segmentation in images aims to train models capable of predicting each pixel's class. Thus, we provide annotated ground truth data that assigns each pixel to the class soil, crop, or weed. Consequently, an approach for this task needs to provide dense predictions assigning each pixel to one of the before-mentioned classes.

State of the Art. Semantic segmentation is a classical task that was first mainly tackled using conditional random fields [40], [43] to exploit the neighboring structure of images. With the advent of deep learning and the success in image classification [41], dense prediction tasks are nowadays mainly tackled by encoder/decoder architectures [54], [76], [77]. Recently, refined architectures add larger context [8], [9] and multi-resolution processing [84] or rely on Transformers [87] for the encoder [12], [97]. We refer to surveys [46], [85] for an overview of recent developments.

In the agricultural domain, most approaches [55], [56], [60] follow the development and adopt the pipelines to account for the row structure [55] or leverage additional background knowledge to cope with less labeled data [60].

Baselines. As baselines, we select DeepLabV3+ [9] (39.8 M params) and ERFNet [76] (2.1 M params) at different ends of model capacity.

Metrics. To evaluate the performance of semantic segmentation models, we report the common intersection-over-union (IoU) for each class individually, where higher values indicate a better performance [14]. Additionally, we compute the mean intersection over union (mIoU) across all classes as the main metric.

Results and Discussion. In Tab. 3, we show quantitative results of the selected baselines. The investigated off-the-shelf semantic segmentation methods already show an overall good performance in terms of mIoU. However, we observe a relatively low IoU for weeds which are often wrongly assigned to pixels of crops. We support these results qualitatively in Fig. 10 and Fig. 11 of the supplement, depicting the predictions of each approach as well as highlighting correct and false predictions. In terms of model capacity, the different investigated methods perform very similarly, indicating that the models' capacity cannot resolve the aforementioned issues. Surprisingly, the smaller, simpler,

Approach	$PQ^{\dagger}$	PQ_{crop}	PQ_{weed}	IoU_{soil}
Panoptic DeepLab [10]	57.97	52.02	22.61	99.27
Mask R-CNN [31]	65.79	67.61	31.30	98.47
Mask2Former [11]	69.99	71.21	40.39	98.38

TABLE 4. Baseline results for panoptic segmentation on the test set.

and faster architecture ERFNet performs on par with the more complex DeepLabV3+ model that commonly shows better performance in the context of autonomous driving. Furthermore, we refer to Tab. 9 of the supplement for more detailed quantitative results distinguishing between each data collection date.

4.2 Panoptic Segmentation

Task description. Panoptic segmentation [38] tackles the task of jointly estimating a pixel-wise semantic label and distinguishing instances. This task differentiates between so-called "stuff" and "thing" classes. The former corresponds to instance-less classes, *i.e.*, soil, and the latter refers to classes with clearly separable objects, *i.e.*, crops and weeds. Consequently, an approach for this task needs to produce semantic masks assigning each pixel to crop, weed, or soil and an instance segmentation for crops and weeds.

State of the Art. Most approaches for panoptic segmentation [37] extend classical semantic segmentation approaches with an instance branch or head to separate "thing" classes. Generally, there are two main paradigms for generating instances prevalent: top-down and bottom-up approaches. Top-down approaches [37], [51], [70] use detection-based bounding box predictions to locate instances and mask predictions in bounding boxes to segment the located instances pioneered by Mask R-CNN [31]. Bottom-up approaches [10], [91] use a separate decoder to estimate embedding vectors and offsets to find clusters corresponding to instances of "thing" classes guided by the semantic segmentation branch. The main focus of research in this field concentrates on improving the architecture to achieve better separation between instances [51], [63], [71]. However, recent approaches [11], [83], [98] based on Vision Transformer [17] show substantial improvements.

In the agricultural domain, most methods adopt panoptic segmentation pipelines for crop and weed detection [6], [28] to contribute towards sustainable crop production and targeted weed management in real field conditions.

Baselines. We use Panoptic DeepLab [10] (7.7 M params) and Mask R-CNN [31] (44.4 M params). Further, we show Mask2Former [11] (44 M params) performance of a Transformer-based approach.

Metrics. We separately compute the panoptic quality [38] for the predicted instance masks of crops (PQ_{crop}) and weeds (PQ_{weeds}). During evaluation, we treat predicted instances associated with a partially visible instance, *i.e.*, a plant where less than 50% of its pixels are inside the image, as "do not care" regions not affecting the score. Additionally, we report the IoU for the semantic segmentation of soil (IoU_{soil}) to consider predictions related to "stuff". In our final metric, we compute the average over all three values and denote it as $PQ^{\dagger}$ as proposed by Porzi *et al*. [70].

Approach	mAP	mAP_{50}	mAP_{75}	AP	
				Crop	Weed
Faster R-CNN [74]	40.43	65.07	40.19	63.23	17.62
Mask R-CNN [31]	38.68	63.72	38.07	60.32	17.05
YOLOv7 [90]	60.48	82.47	62.30	83.06	37.91

TABLE 5. Baseline results for plant detection on the test set.

Results and Discussion. In Tab. 4 we show that Mask2Former [11] achieves the best overall performance. A more detailed quantitative evaluation provided in Tab. 11 of the supplement, distinguishing between different data collection days characterized by specific plant growth stages, reveals that the instance segmentation of plants is challenging in cases of barely visible small plants and large plants with high mutual overlap. We support these results qualitatively in Fig. 13 and Fig. 14 of the supplement. This suggests that domain-specific models could potentially exploit the plant growth stage.

4.3 Detection

Task description. While pixel-wise segmentation of instances allows for extracting fine-grained information, often detecting instances is sufficient. Therefore, we also propose using our data for studying plant or leaf detection in separate tasks. For plant detection, we distinguish between the classes of crop and weed. Similar to COCO [52], we extract bounding box annotations from the instance-level plant and leaf annotations to allow training of object detection approaches. An approach for either plant or leaf detection needs to provide bounding boxes and confidence scores for each detected instance.

State of the Art. Early approaches for object detection relies on sliding window-based classification methods [88] and research before 2014 mainly concentrates on better feature representations [15], [24], part-based representations [23], [50], or better proposal generation [86].

Since 2013, CNN-based approaches have been prevalent as pioneered by R-CNN [26] and follow-up work [25], [31], [74]. Generally, one can distinguish between single-stage and two-stage approaches. Nowadays, single-stage approaches are mainly employed and YOLO [73]-based approaches are popular choices. Recently, also keypoint-based approaches [47], [99] were proposed that divert from the anchor-based methods. Similarly to other tasks, the field recently shifted towards Transformer-based approaches [5].

In the agricultural domain, most methods use detectors to identify crops or weeds [28], [29] or suggest domain-specific adaptations, *e.g.*, for fruit detection [59].

Baselines. We select established approaches for object detection, such as Faster RCNN [74] (41.7 M params), Mask R-CNN [31] (44.4 M params) and YOLOv7 [90] (37.2 M params), which are commonly used approaches. Since this task refers to either plant or leaf detection, we train models for each task separately. Although Mask R-CNN also provides an instance segmentation, we do not consider these here but rely on its predicted bounding boxes.

Metrics. In line with established benchmarks [20], [21], [52], we report the average precision (AP) for each class and mean average precision (mAP) across all classes, which uses

Approach	mAP	mAP_{50}	mAP_{75}
Faster R-CNN [74]	33.91	64.61	31.30
Mask R-CNN [31]	34.41	66.02	32.15
YOLOv7 [90]	57.90	86.85	62.92

TABLE 6. Baseline results for leaf detection on the test set.

Approach	PQ_{leaf}
Mask R-CNN [31]	59.74
Mask2Former [11]	57.50

TABLE 7. Baseline results for leaf instance segmentation on test set.

multiple IoUs for matching between 0.5 and 0.95 with a step size of 0.05. Furthermore, we report the mean average precision at 0.5 IoU (mAP_{50}) and 0.75 IoU (mAP_{75}). As previously, we treat each predicted bounding box associated with a partially visible instance as "do not care" regions. Thus, these predictions do not affect the scores.

Results and Discussion. In Tab. 5, we show results for plant detection, where we see that modern approaches have a clear edge over the other approaches. Apparently, weed detection is more difficult than crop detection, which could result from smaller plant sizes, as also suggested qualitatively in Fig. 16 and Fig. 17 of the supplement.

In Tab. 6, we summarize the results for leaf detection, which shows lower performance across all methods compared with aforementioned plant detection, indicating the need for domain-specific approaches. In Tab. 15 of the supplement, we provide more detailed results for each data collection day and additionally show qualitative results in Fig. 18 and Fig. 19 of the supplement.

4.4 Leaf Instance Segmentation

Task description. Leaf instance segmentation is relevant for estimating the growth stage of a plant [45] and also the basis for leaf disease detection [64]. Such approaches are involved in phenotyping activities to investigate new varieties of crops [62]. An automatic, vision-based assessment of such traits has the potential to have reproducible and objective measurements at a high temporal frequency. Consequently, an approach for this task needs to predict an instance mask for each visible crop leaf.

State of the Art. Instance segmentation is closely related to object detection. Therefore earlier approaches rely on object detection approaches [73], [74] to perform top-down instance segmentation by predicting segmentation masks for bounding boxes [2], [31]. A different line of research [4] investigated the usage of bottom-up processing, where first pixel-wise embedding vectors are estimated such that pixels belonging to the same instance are near in embedding space, while embedding vectors of different instances are separated. The estimated embedding vectors can then be clustered, resulting in instances. Recently, several methods [92], [93] were proposed that directly estimate masks for each object instance. Most recently, also Transformer-based approaches [11], [48] for instance segmentation gained interest. Popularized by CVPPP [61], several approaches tackle the task of leaf instance segmentation [33] or leaf counting [95].

Approach	$PQ^{\dagger}$	PQ	PQ_{crop}	PQ_{leaf}	IoU	
					Weed	Soil
HAPT [75]	65.27	50.73	54.61	46.84	61.11	98.50
Weyler *et al.* [94]	-	40.49	38.37	42.60	-	-

TABLE 8. Baseline results for hierarchical panoptic segmentation on the test set.

Baselines. As baselines for our experiments, we employ Mask R-CNN [31] (44.4 M params) and Mask2Former [11] (44 M params). While the former method represents a traditional top-down approach, the latter belongs to more recent methods relying on a Transformer decoder and masked attention.

Metrics. We compute the panoptic quality [38] for the predicted instance masks of crop leaves, denoted as PQ_{leaf}. As previously, any instance prediction associated with a partially visible instance does not affect the score.

Results and Discussion. Tab. 7 shows the results of the investigated baselines. In this setting, the approaches generally struggle to separate leaves, as they are naturally overlapping, even for smaller plants. In Fig. 20 and Fig. 21 of the supplement, we support these results qualitatively and provide more detailed metrics differentiating between each data collection day in Tab. 17 of the supplement. Again, we suspect that more domain-specific approaches could induce prior knowledge to achieve a better separation.

4.5 Hierarchical Panoptic Segmentation

Task description. Models for hierarchical panoptic segmentation target objects, which can be represented as an aggregation of individual parts, *e.g.*, plants can be represented as the union of their leaves [94]. Consequently, these methods provide a simultaneous instance segmentation of the whole object and each part. Thus, they are capable of providing more detailed information about each object, *e.g.*, the association of individual leaves to a specific plant allows obtaining the total number of leaves per plant, which correlates to its growth stage [45]. We provide the annotated instance masks of all crops and their associated leaves. Since there are no leaf annotations for weeds, we do not consider them under the guise of a hierarchical structure. Thus, we also relate to weeds as "stuff" for this task.

State of the Art. Several recent works exploit the underlying hierarchical structure of objects to obtain a panoptic segmentation [75], [94]. In the agricultural domain, recent methods [75], [94] operating in real field conditions exploit the hierarchical structure of plants to predict the instance segmentation of individual crops and their leaves.

Baselines. We select the methods by Weyler *et al.* [94] (2.2 M params) and Roggiolani *et al.* [75] (2.4 M params) as baselines that both perform a simultaneous instance segmentation of crops and their associated leaves, where the latter method is denoted as HAPT. The first method is a bottom-up approach that first predicts leaves, which are then associated to a plant. In contrast, HAPT uses a hierarchical feature aggregation starting at the plants providing plant-level features to then predict leaves.

Metrics. To evaluate the performance of this task, we compute the panoptic quality [38] for the predicted instance

masks of all crops (PQ_{crop}) and leaves (PQ_{leaf}) separately. We report the average panoptic quality over both values, denoted as PQ. As previously, any instance prediction assigned to a partially visible instance does not affect the metrics. To account for methods that filter pixels related to weeds or soil with an additional semantic segmentation, we also report the IoU for both classes. Finally, we compute $PQ^{\dagger}$ as the average over PQ_{crop}, PQ_{leaf}, and both IoU values.

Results and Discussion. In Tab. 8, we show the results of the hierarchical approaches. Here, we can see that both methods do not obtain consistent predictions for plants at a large growth stage, where individual plants and their leaves overlap. In particular, instance separation of leaves seems most challenging in line with the plant instance segmentation. Thus, methods targeting these scenarios could improve the performance. We support these findings in Tab. 19 of the supplement, where we perform the evaluation for each data collection day separately. Ultimately, we show quantitative results in Fig. 22, which we separate into true positives, false positives, and false negatives in Fig. 23 in the supplement.

5 Challenge in Conjunction with CVPPA Workshop at IEEE/CVF ICCV 2023

In conjunction with the workshop on Computer Vision in Plant Phenotyping and Agriculture held at the IEEE/CVF International Conference on Computer Vision (ICCV) in 2023, we invited the community to tackle the most challenging task of hierarchical panoptic segmentation using our dataset. We received overall 148 submissions from 107 registered participants on the competition hosted on CodaLab[4], where one could upload predictions until a fixed deadline.

For the top-performing entries of the leaderboard, we invited authors to provide a technical report of their approach[5]. The technical solutions surpassed the baselines by a large margin and often employed the Segment Anything Model [39] either in conjunction with a detection approach or initial segmentation that is refined. But also a Mask2Former-based [11] approach using a mask refinement on small plants and a second stage for leaf instance segmentation on plant masks showed promising results surpassing our off-the-shelf baselines presented in Sec. 4.5.

6 Potential Impact on Other Topics

Besides the already covered supervised tasks in agricultural perception, our dataset providing labeled and unlabeled images has the potential to impact also other fields of research and applications in the agricultural domain, such as research in self-supervised representation learning, domain generalization, and unsupervised domain adaptation that is currently getting increasing interest in the computer vision and robotics community. Exploiting developments in semi-supervised, but also unsupervised learning of vision models seems like a indispensable step to reduce the burden of annotating data and unlocking the scalable deployment of vision models in the agricultural domain.

4. The concluded and now closed competition is still available at https://codalab.lisn.upsaclay.fr/competitions/13904.

5. Non-archival, non-peer reviewed technical reports are available at https://cvppa2023.github.io/challenges/

Furthermore, the combination with other agricultural datasets providing pixel-wise annotations, *e.g.*, GrowliFlowers [36], opens the door for studying cross-domain transfer between different plant species towards the goal of developing more generalizable visual perception systems in the agricultural domain.

7 Conclusion

In this paper, we present a novel dataset for studying visual perception in the agricultural domain of crop production using real-world field images captured by an UAV. Together with dense pixel-wise annotations of crops and weeds that distinguish instances of plants, we also provide leaf-level pixel-wise annotations of crop leaves.

In line with the dataset, we presented our benchmark tasks that will be evaluated on a hidden test set to allow an unbiased and controlled evaluation of developed approaches. The server-side evaluation also ensures that metrics are consistent and reliable allowing to compare approaches based on published results.

For each task, we also provide baseline results that show the performance of off-the-shelf approaches for the different tasks. These results show that certain tasks need further research to tackle the specific challenges of the agricultural domain. We believe that more domain-specific approaches exploiting domain knowledge could boost performance.

Acknowledgments

We thank all students annotating the data. The work has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy, EXC-2070 – 390732324 (PhenoRob).

References

[1] J. Bell and H. M. Dee, "Aberystwyth Leaf Evaluation Dataset," https://doi.org/10.5281/zenodo.168158, 2016. 2, 3

[2] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "YOLACT++ Better Real-Time Instance Segmentation," *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 44, no. 2, pp. 1108–1121, 2022. 8

[3] P. Bosilj, E. Aptoula, T. Duckett, and G. Cielniak, "Transfer learning between crop types for semantic segmentation of crops versus weeds in precision agriculture," *Journal of Field Robotics (JFR)*, vol. 37, no. 1, pp. 7–19, 2019. 3

[4] B. D. Brabandere, D. Neven, and L. V. Gool, "Semantic Instance Segmentation with a Discriminative Loss Function," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017. 8

[5] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers," in *Proc. of the Europ. Conf. on Computer Vision (ECCV)*, 2020. 7

[6] J. Champ, A. Mora-Fallas, H. Goëau, E. Mata-Montero, P. Bonnet, and A. Joly, "Instance segmentation for the fine detection of crop and weed plants by precision agricultural robots," *Applications in Plant Sciences*, vol. 8, no. 7, p. e11373, 2020. 7

[7] N. Chebrolu, P. Lottes, A. Schaefer, W. Winterhalter, W. Burgard, and C. Stachniss, "Agricultural Robot Dataset for Plant Classification, Localization and Mapping on Sugar Beet Fields," *Intl. Journal of Robotics Research (IJRR)*, vol. 36, pp. 1045–1052, 2017. 2, 3

[8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic Image Segmentation withDeep Convolutional Nets, Atrous Convolution,and Fully Connected CRFs," *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 40, no. 4, pp. 834–848, 2018. 6

[9] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking Atrous Convolution for Semantic Image Segmentation," *arXiv preprint:1706.05587*, 2017. 6

[10] B. Cheng, M. D. Collins, Y. Zhu, T. Liu, T. S. Huang, H. Adam, and L.-C. Chen, "Panoptic-DeepLab: A Simple, Strong, and Fast Baseline for Bottom-Up Panoptic Segmentation," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 7

[11] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention Mask Transformer for Universal Image Segmentation," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 7, 8, 9

[12] B. Cheng, A. G. Schwing, and A. Kirillov, "Per-Pixel Classification is Not All You Need for Semantic Segmentation," in *Proc. of the Conf. on Neural Information Processing Systems (NeurIPS)*, 2021. 6

[13] M. T. Chiu, X. Xu, Y. Wei, Z. Huang, A. G. Schwing, R. Brunner, H. Khachatrian, H. Karapetyan, I. Dozier, and G. Rose, "Agriculture-vision: A large aerial image database for agricultural pattern analysis," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[14] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes Dataset for Semantic Urban Scene Understanding," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2, 6

[15] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 886–893. 7

[16] E. David, M. Serouart, D. Smith, S. Madec, K. Velumani, S. Liu, X. Wang, F. P. Espinosa, S. Shafiee, I. S. A. Tahir, H. Tsujimoto, S. Nasuda, B. Zheng, N. Kichgessner, H. Aasen, A. Hund, P. Sadhegi-Tehran, K. Nagasawa, G. Ishikawa, S. Dandrifosse, A. Carlier, B. Mercatoris, K. Kuroki, H. Wang, M. Ishii, M. A. Badhon, C. Pozniak, D. S. LeBauer, M. Lilimo, J. Poland, S. Chapman, B. de Solan, F. Baret, I. Stavness, and W. Guo, "Global Wheat Head Dataset 2021: more diversity to improve the benchmarking of wheat head localization methods," *arXiv preprint:2105.07660*, 2021. 3

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *Proc. of the Intl. Conf. on Learning Representations (ICLR)*, 2021. 7

[18] T. Duckett, S. Pearson, S. Blackmore, B. Grieve, W.-H. Chen, G. Cielniak, J. Cleaversmith, J. Dai, S. Davis, C. Fox, P. From, I. Georgilas, R. Gill, I. Gould, M. Hanheide, A. Hunter, F. Iida, L. Mihalyova, S. Nefti-Meziani, G. Neumann, P. Paoletti, T. Pridmore, D. Ross, M. Smith, M. Stoelen, M. Swainson, S. Wane, P. Wilson, I. Wright, and G.-Z. Yang, "Agricultural Robotics: The Future of Robotic Agriculture," *arXiv preprint: 1806.06762*, 2018. 1

[19] H. Dutagaci, P. Rasti, G. Galopin, and D. Rousseau, "Rose-x: an annotated data set for evaluation of 3d plant organ segmentation methods," *Plant Methods*, vol. 16, no. 1, pp. 1–14, 2020. 3

[20] M. Everingham, S. A. Eslami, L. van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes Challenge – a Retrospective," *Intl. Journal of Computer Vision (IJCV)*, vol. 111, no. 1, pp. 98–136, 2015. 2, 7

[21] 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)*, vol. 88, no. 2, pp. 303–338, 2010. 2, 7

[22] M. Fawakherji, C. Potena, A. Pretto, D. D. Bloisi, and D. Nardi, "Multi-Spectral Image Synthesis for Crop/Weed Segmentation in Precision Farming," *Journal on Robotics and Autonomous Systems (RAS)*, vol. 146, p. 103861, 2021. 2, 3

[23] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object Detection with Discriminatively Trained Part-Based Models," *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 32, no. 9, pp. 1627–1645, 2010. 7

[24] J. Gall, A. Yao, N. Razavi, L. V. Gool, and V. Lempitsky, "Hough Forests for Object Detection, Tracking, and Action Recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 33, no. 11, 2011. 7

[25] R. Girshick, "Fast R-CNN," in *Proc. of the IEEE Intl. Conf. on Computer Vision (ICCV)*, 2015. 7

[26] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014. 7

[27] R. Güldenring, R. E. Adersen, and L. Nalpantidis, "Zoom in on the Plant: Fine-grained Analysis of Leaf, Stem and Vein Instances," *IEEE Robotics and Automation Letters (RA-L)*, vol. 9, no. 2, pp. 1588–1595, 2024. 3

[28] M. Halstead, A. Ahmadi, C. Smitt, O. Schmittmann, and C. McCool, "Crop Agnostic Monitoring Driven by Deep Learning," *Frontiers in Plant Science*, vol. 12, 2021. 7

[29] M. S. Hammad, K. K. Velayudhan, J. Potgieter, and K. M. Arif, "Weed identification by single-stage and two-stage neural networks: A study on the impact of image resizers and weights optimization algorithms," *Frontiers in Plant Science*, vol. 13, 2022. 7

[30] S. Haug and J. Ostermann, "A crop/weed field image dataset for the evaluation of computer vision based precision agriculture tasks," in *Proc. of the European Conference on Computer Vision (ECCV) Workshops*, 2015, pp. 105–116. 2, 3

[31] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. of the IEEE Intl. Conf. on Computer Vision (ICCV)*, 2017. 7, 8

[32] L. Horrigan, R. S. Lawrence, and P. Walker, "How sustainable agriculture can address the environmental and human health harms of industrial agriculture," *Environ. Health Perspect.*, vol. 110, no. 5, pp. 445–456, 2002. 1

[33] W. Huang, S. Deng, C. Chen, X. Fu, and Z. Xiong, "Learning to Model Pixel-Embedded Affinity for Homogeneous Instance Segmentation," in *Proc. of the Conf. on Advancements of Artificial Intelligence (AAAI)*, 2022. 8

[34] M. Imperoli, C. Potena, D. Nardi, G. Grisetti, and A. Pretto, "An Effective Multi-Cue Positioning System for Agricultural Robotics," *IEEE Robotics and Automation Letters (RA-L)*, vol. 3, no. 4, pp. 3685–3692, 2018. 3

[35] R. Khanna, L. Schmid, A. Walter, J. Nieto, R. Siegwart, and F. Liebisch, "A spatio temporal spectral framework for plant stress phenotyping," *Plant Methods*, vol. 15, no. 1, pp. 1–18, 2019. 3

[36] J. Kierdorf, L. V. Junker-Frohn, M. Delaney, M. D. Olave, A. Burkart, H. Jaenicke, O. Muller, U. Rascher, and R. Roscher, "GrowliFlower: An image time series dataset for GROWth analysis of cauLIFLOWER," *Journal of Field Robotics (JFR)*, vol. 40, no. 2, pp. 173–192, 2022. 3, 4, 9

[37] A. Kirillov, R. Girshick, K. He, and P. Dollar, "Panoptic Feature Pyramid Networks," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 7

[38] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, "Panoptic Segmentation," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 7, 8

[39] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick, "Segment anything," in *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision (ICCV)*, 2023. 3, 9

[40] P. Krähenbühl and V. Koltun, "Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials," in *Proc. of the Conf. on Neural Information Processing Systems (NIPS)*, 2011. 6

[41] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017. 1, 6

[42] H. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955. 5

[43] L. Ladicky, C. Russell, and P. Kohli, "Associative Hierarchical CRFs for Object Class Image Segmentation," in *Proc. of the IEEE Intl. Conf. on Computer Vision (ICCV)*, 2009. 6

[44] P. Lameski, E. Zdraveski, V. Trajkovik, and A. Kulkov, "Weed Detection Dataset with RGB Images Taken Under Variable Light Conditions," in *Proc. of the Intl. Conf. on ICT Innovations*, 2017. 2, 3

[45] P. D. Lancashire, H. Bleiholder, T. Boom, P. Langelüddeke, R. Stauss, E. Weber, and A. Witzenberger, "A Uniform Decimal Code for Growth Stages of Crops and Weeds," *Annals of Applied Biology*, vol. 119, no. 3, pp. 561–601, 1991. 2, 8

[46] F. Lateef and Y. Ruichek, "Survey on semantic segmentation using deep learning techniques," *Neurocomputing*, vol. 338, pp. 321–348, 2019. 6

[47] H. Law and J. Deng, "CornerNet: Detecting Objects as Paired Keypoints," in *Proc. of the Europ. Conf. on Computer Vision (ECCV)*, 2018. 7

[48] J. Lazarow, W. Xu, and Z. Tu, "Instance Segmentation with Mask-supervised Polygonal Boundary Transformers," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 8

[49] Y. LeCun, Y. Bengio, and G. E. Hinton, "Deep Learning," *Nature*, vol. 521, pp. 436–444, 2015. 1, 2

[50] B. Leibe, A. Leonardis, and B. Schiele, "Combined Object Categorization and Segmentation with an Implicit Shape Model," in *Proc. of Workshop on Statistical Learning in Computer Vision at ECCV*, 2004. 7

[51] Y. Li, H. Zhao, X. Qi, L. Wang, Z. Li, J. Sun, and J. Jia, "Fully Convolutional Networks for Panoptic Segmentation," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 7

[52] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *Proc. of the Europ. Conf. on Computer Vision (ECCV)*, 2014, pp. 740–755. 1, 2, 7

[53] M. T. Linaza, J. Posada, J. Bund, P. Eisert, M. Quartulli, J. Döllner, A. Pagani, I. G. Olaizola, A. Barriguinha, T. Moysiadis, and L. Lucat, "Data-Driven Artificial Intelligence Applications for Sustainable Precision Agriculture," *Agronomy*, vol. 11, no. 6, p. 1227, 2021. 1

[54] 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. 6

[55] 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)*, vol. 3, pp. 3097–3104, 2018. 6

[56] P. Lottes, M. Höferlin, S. Sander, and C. Stachniss, "Effective Vision-based Classification for Separating Sugar Beets and Weeds for Precision Farming," *Journal of Field Robotics (JFR)*, vol. 34, pp. 1160–1178, 2017. 6

[57] Y. Lu and S. Young, "A survey of public datasets for computer vision tasks in precision agriculture," *Computers and Electronics in Agriculture*, vol. 178, p. 105760, 2020. 2

[58] S. L. Madsen, S. K. Mathiassen, M. Dyrmann, M. S. Laursen, L.-C. Paz, and R. N. Jørgensen, "Open Plant Phenotype Database of Common Weeds in Denmark," *Remote Sensing*, vol. 12, no. 8, p. 1246, 2020. 3

[59] X. Mai, H. Zhang, and M. Q. Meng, "Faster R-CNN with Classifier Fusion for Small Fruit Detection," in *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2018. 7

[60] A. Milioto, P. Lottes, and C. 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. 6

[61] M. Minervini, A. Fischbach, H. Scharr, and S. A. Tsaftaris, "Finely-grained annotated datasets for image-based plant phenotyping," *Pattern Recognition Letters*, vol. 81, pp. 80–89, 2016. 2, 3, 8

[62] M. Minervini, H. Scharr, and S. A. Tsaftaris, "Image Analysis: The New Bottleneck in Plant Phenotyping," *IEEE Signal Processing Magazine*, vol. 32, no. 4, pp. 126–131, 2015. 2, 8

[63] R. Mohan and A. Valada, "EfficientPS: Efficient Panoptic Segmentation," *Intl. Journal of Computer Vision (IJCV)*, vol. 129, pp. 1551–1579, 2021. 7

[64] S. P. Mohanty, D. P. Hughes, and M. Salathé, "Using deep learning for image-based plant disease detection," *Frontiers in Plant Science*, vol. 7, p. 1419, 2016. 8

[65] A. K. Mortensen, S. Skovsen, H. Karstoft, and R. Gislum, "The Oil Radish Growth Dataset for Semantic Segmentation and Yield Estimation," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019. 3

[66] G. Neuhold, T. Ollmann, S. R. Bulo, and P. Kontschieder, "The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes," in *Proc. of the IEEE Intl. Conf. on Computer Vision (ICCV)*, 2017. 1, 2

[67] A. Olsen, D. A. Konovalov, B. Philippa, P. Ridd, J. C. Wood, J. Johns, W. Banks, B. Girgenti, O. Kenny, J. Whinney, B. Calvert, M. R. Azghadi, and R. D. White, "DeepWeeds: A Multiclass Weed Species Image Dataset for Deep Learning," *Scientifc Reports*, vol. 9, no. 1, p. 2058, 2019. 3

[68] A. Pavao, I. Guyon, A.-C. Letournel, D.-T. Tran, X. Baro, H. J. Escalante, S. Escalera, T. Thomas, and Z. Xu, "CodaLab Competitions: An Open Source Platform to Organize Scientific Challenges," *Journal on Machine Learning Research (JMLR)*, vol. 24, no. 198, pp. 1–6, 2023. 3, 5

[69] T. Pire, M. Mujica, J. Civera, and E. Kofman, "The Rosario dataset: Multisensor data for localization and mapping in agricultural environments," *Intl. Journal of Robotics Research (IJRR)*, vol. 38, no. 6, 2019. 3

[70] L. Porzi, S. R. Bulo, A. Colovic, and P. Kontschieder, "Seamless Scene Segmentation," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 7

[71] L. Porzi, S. R. Bulo, and P. Kontschieder, "Improving Panoptic Segmentation at All Scales ," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 7

[72] M. P. Pound, J. A. Atkinson, A. J. Townsend, M. H. Wilson, M. Griffiths, A. S. Jackson, A. Bulat, G. Tzimiropoulos, D. M. Wells, E. H. Murchie, T. P. Pridmore, and A. P. French, "Deep machine learning provides state-of-the-art performance in image-based plant phenotyping," *Gigascience*, vol. 6, no. 10, p. gix083, 2017. 1

[73] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016. 7, 8

[74] 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 Conf. on Neural Information Processing Systems (NIPS)*, 2015. 7, 8

[75] G. Roggiolani, M. Sodano, T. Guadagnino, F. Magistri, J. Behley, and C. Stachniss, "Hierarchical Approach for Joint Semantic, Plant Instance, and Leaf Instance Segmentation in the Agricultural Domain," in *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2023. 8

[76] E. Romera, J. M. Alvarez, L. M. Bergasa, and R. Arroyo, "Erfnet: Efficient residual factorized convnet for real-time semantic segmentation," *IEEE Trans. on Intelligent Transportation Systems (ITS)*, vol. 19, no. 1, pp. 263–272, 2017. 6

[77] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Proc. of the Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015. 6

[78] I. Sa, Z. Ge, F. Dayoub, B. Upcroft, T. Perez, and C. McCool, "DeepFruits: A Fruit Detection System Using Deep Neural Networks," *Sensors*, vol. 16, no. 8, p. 1222, 2016. 3

[79] I. Sa, M. Popovic, R. Khanna, Z. Chen, P. Lottes, F. Liebisch, J. Nieto, C. Stachniss, A. Walter, and R. Siegwart, "WeedMap: A Large-Scale Semantic Weed Mapping Framework Using Aerial Multispectral Imaging and Deep Neural Network for Precision Farming," *Remote Sensing*, vol. 10, no. 9, p. 1423, 2018. 2, 3

[80] D. Schunck, F. Magistri, R. A. Rosu, A. Cornelißen, N. Chebrolu, S. Paulus, J. Léon, S. Behnke, C. Stachniss, H. Kuhlmann, and L. Klingbeil, "Pheno4D: A spatio-temporal dataset of maize and tomato plant point clouds for phenotyping and advanced plant analysis ," *PLOS ONE*, vol. 16, no. 8, pp. 1–18, 2021. 3

[81] D. Steininger, A. Trondl, G. Croonen, J. Simon, and V. Widhalm, "The cropandweed dataset: A multi-modal learning approach for efficient crop and weed manipulation," in *Proc. of the IEEE Winter Conf. on Applications of Computer Vision (WACV)*, 2023. 3

[82] H. Storm, S. Seidel, L. Klingbeil, F. Ewert, H. Vereecken, W. Amelung, S. Behnke, M. Bennewitz, J. Börner, T. Döring, J. Gall, A.-K. Mahlein, C. McCool, U. Rascher, S. Wrobel, A. Schnepf, C. Stachniss, and H. Kuhlmann, "Research Priorities to Leverage Smart Digital Technologies for Sustainable Crop Production," *European Journal of Agronomy*, vol. 156, p. 127178, 2024. 1

[83] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for Semantic Segmentation," in *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision (ICCV)*, 2020. 7

[84] K. Sun, Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, Y. Mu, X. Wang, W. Liu, and J. Wang, "High-Resolution Representations for Labeling Pixels and Regions," *arXiv preprint:1904.04514*, 2019. 6

[85] S. A. Taghanaki, K. Abhishek, J. P. Cohen, J. Cohen-Adad, and G. Hamarneh, "Deep semantic segmentation of natural and medical images: a review," *Artificial Intelligence Review*, vol. 54, no. 1, pp. 137–178, 2021. 6

[86] K. E. A. van de Sande, J. R. R. Uijlings, T. Gevers, and A. W. M. Smeulders, "Segmentation As Selective Search for Object Recognition," in *Proc. of the IEEE Intl. Conf. on Computer Vision (ICCV)*, 2011. 7

[87] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. of the Conf. on Neural Information Processing Systems (NeurIPS)*, 2017. 6

[88] P. Viola and M. J. Jones, "Robust Real-time Object Detection," *Intl. Journal of Computer Vision (IJCV)*, vol. 57, pp. 137–154, 2001. 7

[89] A. Walter, R. Khanna, P. Lottes, C. Stachniss, R. Siegwart, J. Nieto, and F. Liebisch, "Flourish - a robotic approach for automation in crop management," in *Proc. of the Intl. Conf. on Precision Agriculture*, 2018. 1

[90] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," *arXiv preprint: 2207.02696*, 2022. 7, 8

[91] H. Wang, R. Luo, M. Maire, and G. Shakhnarovich, "Pixel Consensus Voting for Panoptic Segmentation," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 7

[92] X. Wang, T. Kong, C. Shen, Y. Jiang, and L. Li, "SOLO: Segmenting Objects by Locations," in *Proc. of the Europ. Conf. on Computer Vision (ECCV)*, 2020. 8

[93] X. Wang, R. Zhang, T. Kong, L. Li, and C. Shen, "SOLOv2: Dynamic and Fast Instance Segmentation," in *Proc. of the Conf. on Neural Information Processing Systems (NeurIPS)*, 2020. 8

[94] J. Weyler, F. Magistri, P. Seitz, J. Behley, and C. Stachniss, "In-Field Phenotyping Based on Crop Leaf and Plant Instance Segmentation," in *Proc. of the IEEE Winter Conf. on Applications of Computer Vision (WACV)*, 2022. 8

[95] J. Weyler, A. Milioto, T. Falck, J. Behley, and C. Stachniss, "Joint Plant Instance Detection and Leaf Count Estimation for In-Field Plant Phenotyping," *IEEE Robotics and Automation Letters (RA-L)*, vol. 6, no. 2, pp. 3599–3606, 2021. 8

[96] W. Winterhalter, F. V. Fleckenstein, C. Dornhege, and W. Burgard, "Crop Row Detection on Tiny Plants With the Pattern Hough Transform," *IEEE Robotics and Automation Letters (RA-L)*, vol. 3, no. 4, pp. 3394–3401, 2018. 3

[97] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers," in *Proc. of the Conf. on Neural Information Processing Systems (NeurIPS)*, 2021. 6

[98] Q. Yu, H. Wang, D. Kim, S. Qiao, M. Collins, Y. Zhu, H. Adam, A. Yuille, and L.-C. Chen, "CMT-DeepLab: Clustering Mask Transformers for Panoptic Segmentation ," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 7

[99] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as Points," *arXiv preprint:1904.07850v2*, 2019. 7

Jan Weyler is a PhD student in Engineering at the Photogrammetry & Robotics Lab at the University of Bonn, Germany. He obtained his B.Sc. in 2015 and his M.Sc. degree in Geodesy and Geoinformation in 2019 from the University of Bonn, Germany. His research focuses on vision-based semantic scene understanding for agricultural robots.

Federico Magistri is a Ph.D. student at the Photogrammetry & Robotics Lab at the University of Bonn, Germany, since November 2019. He received his M.Sc. in Artificial Intelligence and Robotics from "La Sapienza" University of Rome, Italy, with a thesis on Swarm Robotics for Precision Agriculture in collaboration with the National Research Council of Italy and the Wageningen University and Research, Netherlands.

Elias Marks is a PhD student in Engineering at the Photogrammetry & Robotics Lab at the University of Bonn, Germany. He obtained his B.Sc. degree in Robotics and Automation from the Hochschule Heilbronn, Germany, in 2018 and received his M.Sc. degree in Artificial Intelligence and Robotics at University La Sapienza in Rome, Italy, in 2021. His research focuses on plant modeling for phenotyping based on image data.

Yue Linn Chong is a Ph.D. student in Engineering at Photogrammetry & Robotics Lab at the University of Bonn, Germany. She completed her B.Eng in Mechanical Engineering from the National University of Singapore in 2017. In 2020, she completed her M.Sc. in Mechanical Engineering from the National University of Singapore. Her research focuses on unsupervised learning using generative models.

Matteo Sodano is a PhD student in Engineering at the Photogrammetry & Robotics Lab at the University of Bonn since January 2021. He obtained his MSc degree in Control Engineering in 2020. His research centers around perception and segmentation, with a focus on novel object discovery.

Gianmarco Roggiolani is a Ph.D. candidate in the Photogrammetry & Robotics Lab at the University of Bonn, Germany. He obtained his B.Sc. degree in Computer and Automatic Engineering in 2018 and received his MSc degree in Artificial Intelligence and Robotics in 2021, both from the Sapienza University of Rome, Italy. His research focuses on self-supervised techniques to improve the performance of vision-based learning systems in agricultural robotics.

Nived Chebrolu is a postdoctoral research associate at the Oxford Robotics Institute, University of Oxford, UK. His research interests are in developing robust localization and mapping techniques for field robotics applications. He obtained his Ph.D. from the University of Bonn in 2021, where he developed registration techniques for agricultural robotic applications. Before that, Nived received his M.Sc. in Robotics from Ecole Centrale de Nantes (ECN), France, and the University of Genoa, Italy in 2015.

Cyrill Stachniss is a full professor at the University of Bonn, Germany, with the University of Oxford, UK, as well as with the Lamarr Institute for Machine Learning and AI, Germany. He is the Spokesperson of the DFG Cluster of Excellence PhenoRob at the University of Bonn. His research focuses on probabilistic techniques and learning approaches for mobile robotics, perception, and navigation. Main application areas of his research are agricultural and service robotics and self-driving cars.

Jens Behley received his Dipl.-Inform. in computer science in 2009 and his Ph.D. in computer science in 2014, both from the Dept. of Computer Science at the University of Bonn, Germany. Since 2016, he is a postdoctoral researcher at the Photogrammetry & Robotics Lab at the University of Bonn, Germany. He finished his habilitation at the University of Bonn in 2023. His area of interest lies in the area of perception for autonomous vehicles, deep learning for semantic interpretation, and LiDAR-based SLAM.

Supplementary Material

PhenoBench — A Large Dataset and Benchmarks for Semantic Interpretation in the Agricultural Domain

1 Data Collection

As mentioned in the paper, we performed multiple missions using the same sensor setup during 2020 and 2021 on different fields to collect our dataset. We emphasize that the training and validation set contains only images captured in 2020, while the test set includes data from both years. Thus, any model achieving high performance on the test set must generalize to different fields captured across multiple years, an essential property for real-world applications. In Fig. 6 and Fig. 7, we show images from the test sets captured in 2020 and 2021 to emphasize the variation in lighting conditions and changes in the visual appearance of plants.

2 Quality Assurance Process

As explained in the main paper, we followed a rigorous validation process to ensure high-quality and accurate annotations. See our website, www.phenobench.org, for some qualitative examples of the annotations.

For hiring students for annotation, we employed an initial task of annotating a single image to ensure that students could achieve the required annotation quality and can spot crops and weeds correctly. The selection process helped us to ensure that we have student annotators that provide a good quality level of annotations.

To have consistent annotations of plant and leaves, we split the annotations in two phases: (1) plant mask annotation and (2) leaf mask annotation, where we used the plant masks to limit the leaf labels, as we want to ensure that leaf labels only appear where are also plant masks annotated.

Each phase was closely supervised by a team of researchers, which we call now senior annotators, *i.e.*, PhD students and post docs working in the domain of agricultural computer vision. All researchers published in the agricultural domain and worked with sugar beet data before. Each of these senior annotators supervised 2-3 student annotators, *i.e.*, students with different background including biology, plant sciences, and computer science-related fields of study. The senior annotators provided guidance and feedback after each iteration. After finishing a batch of images by the student annotators, *i.e.*, all four iterations, the senior annotators did a final pass of corrections to ensure that plant annotations and leafs are consistent and accurate.

After all individual annotations were completed by a team of student and senior annotators, we had a final round of corrections, where each batch of images were corrected and approved by another senior annotator including corrections and clean-up of annotations. By having another round of corrections with a different senior annotator, we ensure that the quality of the annotations achieves a high standard and remaining errors that slipped through in the first round of corrections get removed. Note that we then corrected the global images directly instead of following the iterative annotation process described in the main paper.

For validation of the annotations, we also improved our custom annotation tool (see Sec. 4) to make the process of corrections and validation easier by providing ways of iterating through plant instances, which ensured that the senior annotator checks each crop and weed.

A common issue that we noticed in the final round of correction of the annotations, where inaccuracies in the plant boundaries caused by shadows or texture of the ground that was wrongly identified as dry parts of the leaves. Annotations of images recorded on days with overcast weather conditions were generally better annotated as the difference between plant and background was more clearly visible.

Furthermore, we observed that the annotators or senior annotators improved substantially over time in respect to the quality of annotations. In particular, the identification of leaves got easier over time as the general appearance and possible arrangements of leaves got clearer over time. Due to these improvements in annotation quality, we believe that the second and final round of annotations substantially improved the consistency and accuracy of the annotations.

To ensure consistency between plant and leaf annotations, we also had a last algorithmic consistency check that removed leaf annotations that are outside of a plant annotation, which could be caused by the final validation round where we corrected semantic masks of plants.

3 Development Kit

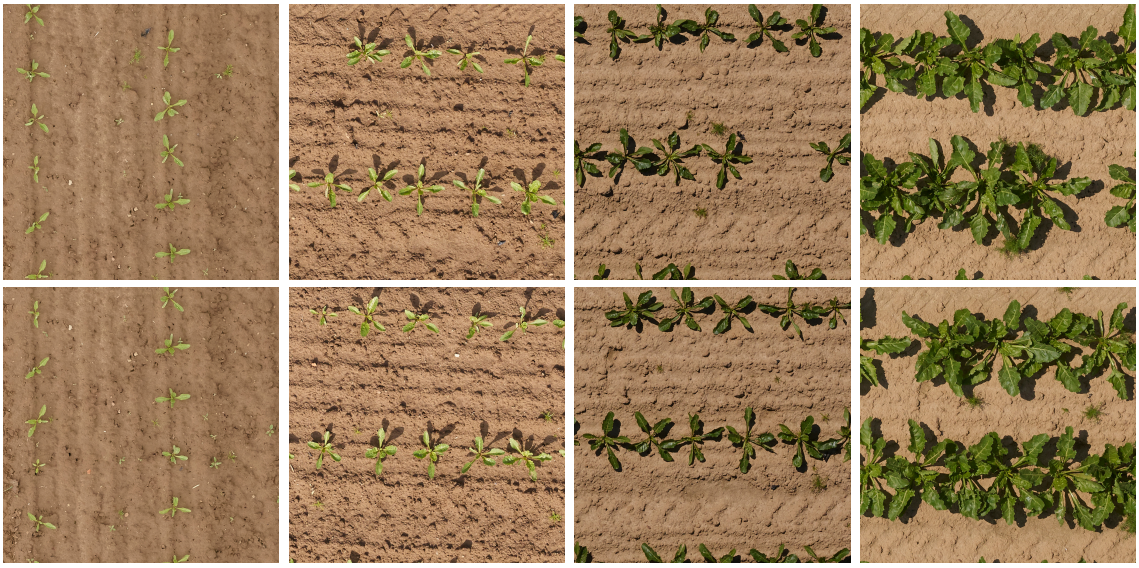
For simple data access and reproducibility of the employed task-specific metrics, we also provide a devkit that includes a PyTorch [15] dataloader, evaluation scripts for computing the metrics, and scripts for visualization of the results publicly available at https://github.com/PRBonn/phenobench.

4 Image Labeler

In this section, we provide additional details on the annotation tool we used to annotate the images. We implemented a tool for our specific needs since publicly available tools did not provide the required capabilities to ensure a consistent hierarchical annotation. Furthermore, it allowed us to modify the tool for our specific needs. To this end, we implemented a Python-based application.

(a) 2020/05/15 (b) 2020/05/26 (c) 2020/06/05

Fig. 6. Sample images from the test set captured at different dates in 2020.



(a) 2021/05/20 (b) 2021/05/28 (c) 2021/06/01 (d) 2021/06/10

Fig. 7. Sample images from the test set captured at different dates in 2021.

Fig. 8 shows an overview of our tool, where we indicated different sections and capabilities via the numbers on the left side.

In the following, we shortly summarize the capabilities of our so-called "image labeler" and the design decisions. We plan to release also our annotation tool since we believe that our modular design and distinct capabilities make it a valuable contribution for the community.

4.1 Layers

As we intended to provide a hierarchical annotation, where leaves should always be entirely inside the corresponding plant mask, we had to represent the different levels of masks in the labeler.

For this purpose, we used the folder structure of the annotations to denote the parent-child relations. A layer is specified by a file called `manifest.yaml` inside the folders, such that the type of the layer and the available set of categories are specified. Categories that can be selected for the currently active layer (see "category selection" (6) in Fig. 8). An annotation layer is selected via the list field shown under "layer selection" (1) in Fig. 8. Depending on the selected layer and type, the categories under "category selection" (6) and the available tools under "toolbar" (5) change.

Currently, we have a "semantic" layer implemented that can be used for pixel-wise semantic annotations. However, other types of annotation layers, such as bounding boxes, key points, etc., should be possible to be integrated into our annotation tool.

4.2 Image Enhancement

We provide some common image enhancement capabilities (see "image enhancement" (2) in Fig. 8) to change the brightness, contrast, and gamma of the image. This helps, especially when the lighting conditions make it hard to identify leaves in the shadow areas caused by the self-occlusions of the plant.

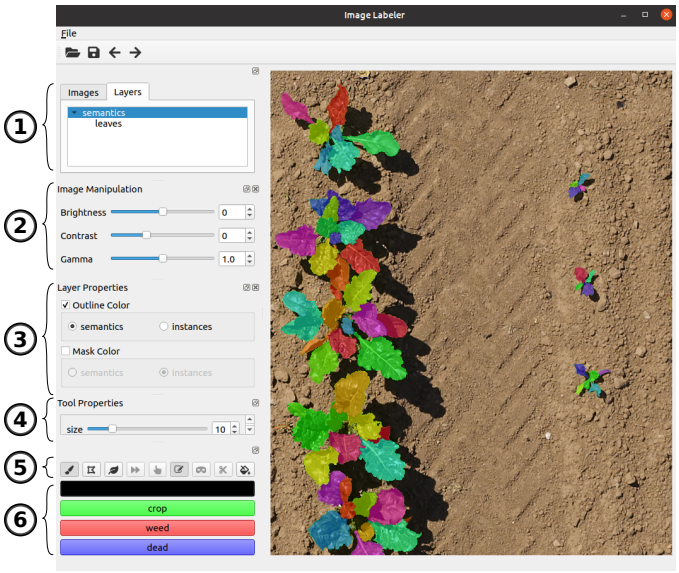


Fig. 8. Our "image labeler" for annotating plant images implemented in Python with OpenGL-based drawing capabilities. On the left, we indicate the different sections of our tool: (1) layer selection, (2) image enhancement, (3) layer properties, (4) tool properties, (5) a toolbar, and (6) category selection.

4.3 Layer Properties

We allow the user also to select, which information is visualized inside the image, *i.e.*, one can switch between semantic and instance colors and select between outlines and masks. These properties are selectable for each layer to select the most relevant information for the annotation. For instance, when annotating leaves, we usually show just the plants' semantic outline and the instance masks of leaves to better distinguish between different leaves.

4.4 Tool properties

Depending on the selected tool, we also can show different properties, *e.g.*, brush size, as shown under "tool properties" (4) in Fig. 8. The properties are specific to each tool. We provide properties for boolean, float that are translated into a widget, such as a slider or a checkbox.

4.5 Tools

As mentioned before, we have layer-specific tools, where we show here the tools that we implemented for the dense pixel-wise semantics:

- A brush () to annotate pixels with a resizable paint brush using the currently active label.
- The polygon tool () allows to specify areas by setting the vertices of a polygon that are then filled with the currently selected label.
- The vegetation mask () allows to show an excess-green-based vegetation mask that can be adjusted via the tool parameters. As with the masking (), we label only pixels that are highlighted via the mask of the vegetation mask.
- As we require unique instance ids, we have a dedicated tool () to advance the instance id value. Instance ids are unique within a given image.
- To select a specific instance, which is often needed to refine labels, we provide an instance selection tool (), which allows selecting an instance of the currently active category.
- We also provide a way to toggle the overwriting of existing labels ().
- For the hierarchical annotation, we provide a masking tool (), which applies a mask following the corresponding instance on the parent layer. The selected parent mask will then restrict all annotations done via the brush and polygon to the area of the parent mask.
- During annotation, we found that it was a common mistake to accidentally label multiple plants as a single instance. To correct such cases, we implemented a cutting tool () to separate an instance into multiple parts.
- Lastly, we provide a fill tool () that allows to re-label or add regions to an existing instance label.

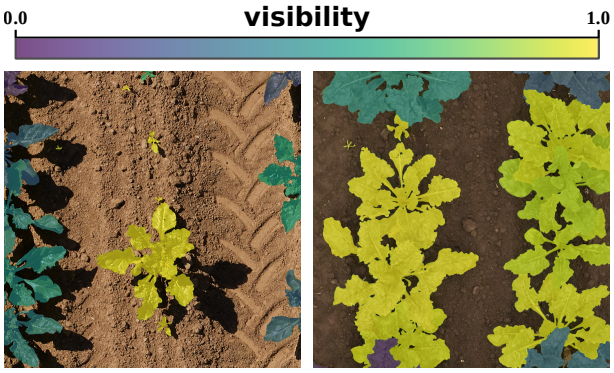


Fig. 9. Examples of visibility plant masks. We indicate the amount of visible pixels from the complete annotation by the color ranging from dark blue to yellow. Plants below a visibility of 0.5 are treated as partially visible.

Date	Approach	mIoU	IoU		
			Crop	Weed	Soil
05-15-2020	ERFNet [18]	81.47	91.74	53.02	99.65
	DeepLabV3+ [1]	81.02	91.28	52.14	99.63
05-26-2020	ERFNet [18]	84.98	94.64	61.49	98.82
	DeepLabV3+ [1]	84.95	94.48	61.62	98.75
06-05-2020	ERFNet [18]	91.31	96.83	78.49	98.62
	DeepLabV3+ [1]	91.29	96.83	78.40	98.64
2020	ERFNet [18]	86.70	94.46	66.28	99.36
	DeepLabV3+ [1]	86.56	94.25	66.08	99.33
05-20-2021	ERFNet [18]	82.08	90.82	55.87	99.53
	DeepLabV3+ [1]	80.97	89.14	54.27	99.49
05-28-2021	ERFNet [18]	87.03	92.81	68.82	99.46
	DeepLabV3+ [1]	87.38	92.54	70.14	99.45
06-01-2021	ERFNet [18]	77.92	91.16	43.67	98.94
	DeepLabV3+ [1]	78.72	89.22	48.19	98.75
06-10-2021	ERFNet [18]	78.53	93.91	44.38	97.30
	DeepLabV3+ [1]	80.65	93.75	51.07	97.14
2021	ERFNet [18]	80.11	93.61	48.35	98.37
	DeepLabV3+ [1]	81.38	93.28	52.62	98.26

TABLE 9. Baseline results for semantic segmentation on the test set, separated to 2020, 2021, and each data collection day.

Date	Approach	mIoU	IoU		
			Crop	Weed	Soil
05-15-2020	ERFNet [18]	80.43	91.90	49.75	99.64
	DeepLabV3+ [1]	80.15	91.38	49.45	99.62
05-26-2020	ERFNet [18]	84.28	94.88	58.68	99.26
	DeepLabV3+ [1]	84.16	94.73	58.51	99.23
06-05-2020	ERFNet [18]	90.29	96.46	75.82	98.60
	DeepLabV3+ [1]	89.91	96.33	74.85	98.56
2020	ERFNet [18]	87.77	95.34	68.65	99.33
	DeepLabV3+ [1]	87.47	95.13	68.00	99.30

TABLE 10. Baseline results for semantic segmentation on the validation set recorded at different data collection days in 2020.

5 Visibility Masks

As mentioned in the paper, we annotated complete plants enabled by the overlapping tiling, ensuring that plants are at least in one of the iterations completely visible. Therefore, we can also account for the amount of visible pixels in the evaluation and provide this information in the training and validation set. Thus, it is possible to account for the visibility of plants when training and developing the approaches. We provide the same information for individual crop leaf instances as well.

Fig. 9 shows examples of the provided visibility masks, where the colors indicate the percentage of visible pixels inside the image of the complete annotated plant.

6 Baselines

In this section, we provide additional information on the baselines, including the training setup, results on the validation set, and qualitative results comparing the different approaches. We separate the baselines by each benchmark task described in the paper.

To ensure reproducibility of the baselines, we additionally also provide implementations with configs, checkpoints, and results at https://github.com/PRBonn/phenobench-baselines. For most baselines, we provide docker containers to ensure that the provided code can be run on various systems where the docker platform is available. Each baseline directory also contains the baseline-specific configuration files, and we summarize here the most important hyperparameters of the training setup to provide an overview.

We additionally provide validation set results to enable comparison of novel approaches in ablation studies using the validation set. The provided results also suggest that the validation set performance is a good indicator of test set performance, which is achieved by having a sufficiently large validation set.

For the qualitative results, we show images of different dates and, consequently, different growth stages. As described in the paper, earlier dates correspond to early growth stages where crops are clearly separated and usually show only a few well-separated leaves. At the last date, the plants show a significantly larger overlap and a larger number of leaves. While early growth stages seem to be less of an issue, the later growth states show a substantial drop in performance, particularly in the panoptic segmentation and hierarchical panoptic segmentation tasks. In these late growth stages, we see the most need for further research.

All approaches were trained on a single Nvidia RTX A6000 with 48 GB of memory using PyTorch [15], where we used the version required by each baseline implementation respectively.

6.1 Semantic Segmentation

We employ ERFNet [18] and DeepLabV3+ [1] based on ResNet-50 [9] as architectures for semantic segmentation and provide here more details on these baseline approaches.

Training setup. We use the same training setup for both approaches. Specifically, we train each model for 4096 epochs with a batch size of 4 using a weighted cross-entropy loss [11]. During optimization, we employ Adam [12] and set the weight decay to $2 \cdot 10^{-4}$. At the initial 16 epochs, we linearly increase the learning rate to $1 \cdot 10^{-4}$ and subsequently apply a polynomial learning rate decay $\left(1 - \frac{e}{4096}\right)^3$, where e is the current epoch. During training, we apply standard data augmentation methods to the input image, *i.e.*, random adjustment of brightness, contrast, hue, and saturation, as well as random scaling and horizontal or vertical flipping. Furthermore, we feed randomly cropped patches of size 768 px × 768 px from the input image to the network.

Evaluation on the 2020 and 2021 Test Set. Tab. 9 provides quantitative results on the 2020 and 2021 data, belonging to the test set. Since we train the models only on 2020 data, both ERFNet and DeepLabV3+ suffer a drop in performance when we test on 2021 data.

Evaluation on the Validation Set. Tab. 10 provides quantitative results on the validation set. The validation set numbers are apparently good indicators of test set performance and, therefore, should ensure that insight on the validation set transfers well to the test set, which seems important when we restrict the number of test set submissions.

Qualitative Results. In Fig. 10, we show qualitative results for ERFNet [18] and DeepLabV3+ [1]. We encode the predictions of each class by a specific color, *i.e.*, crops in green and weeds in red. As the quantitative results suggest, both networks often assign pixels to crops that actually belong to weeds. Contrary, most pixels annotated as crop in the ground truth are correctly predicted by both approaches. Furthermore, we highlight in Fig. 11 each mispredicted pixel in magenta, i.e., where the prediction does not match the ground truth class. Here, it becomes also evident that many mispredictions occur at the contour of plants. We trace this back to label noise in these regions, as such pixels are particularly difficult to assign to the background or vegetation.

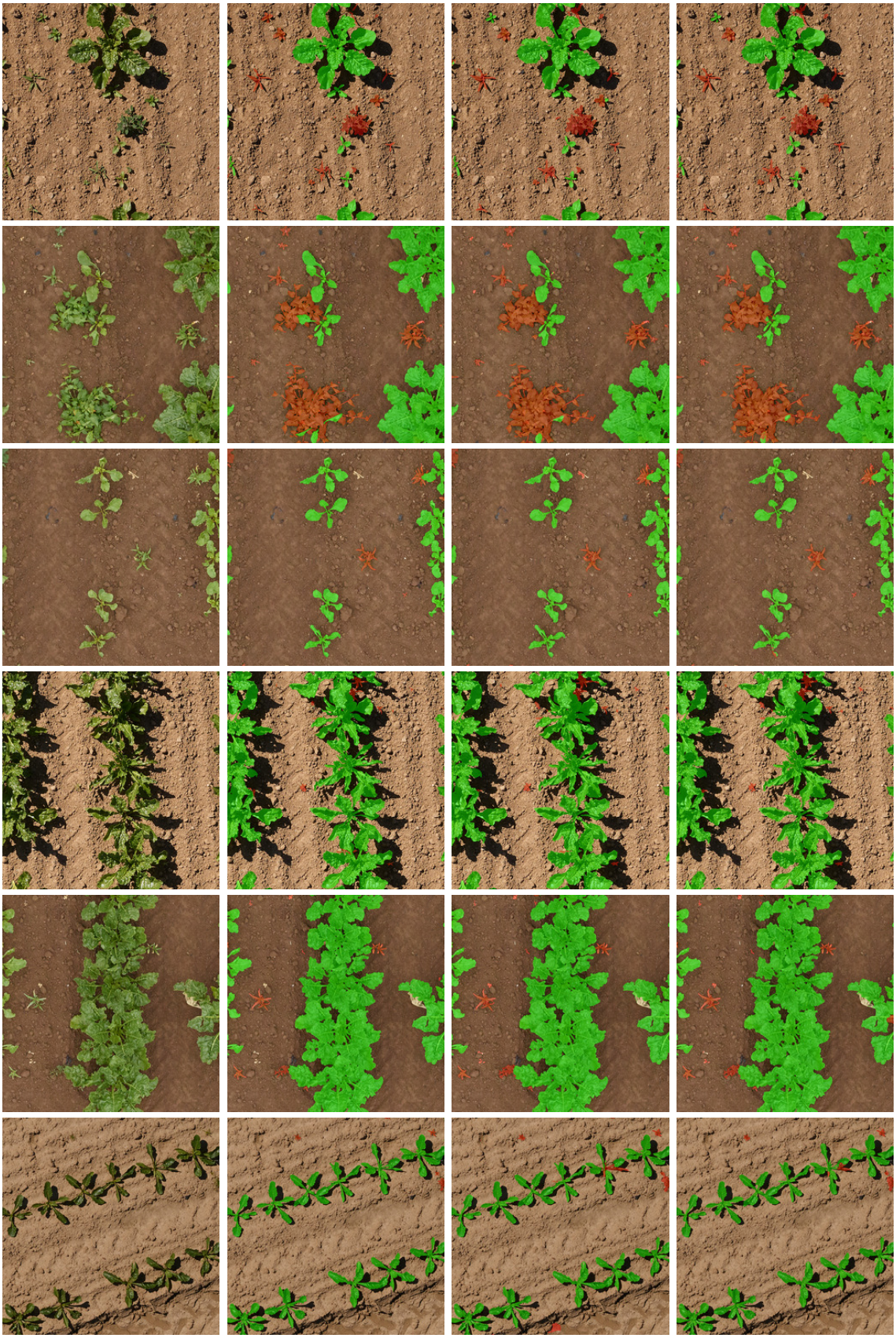
Overall, the predicted semantic segmentation quality is already at a satisfactory level, but this is an expected outcome as the task is basically characterized by predicting a vegetation mask and then assigning a class to individual pixels. As we show later, predicting instance masks and bounding boxes of plants is a much more challenging task since these approaches also need to distinguish between different crops with substantial overlap between plants in later growth stages.

6.2 Panoptic Segmentation

We investigate three commonly employed methods for panoptic segmentation and provide more details about their training procedures in the following paragraphs.

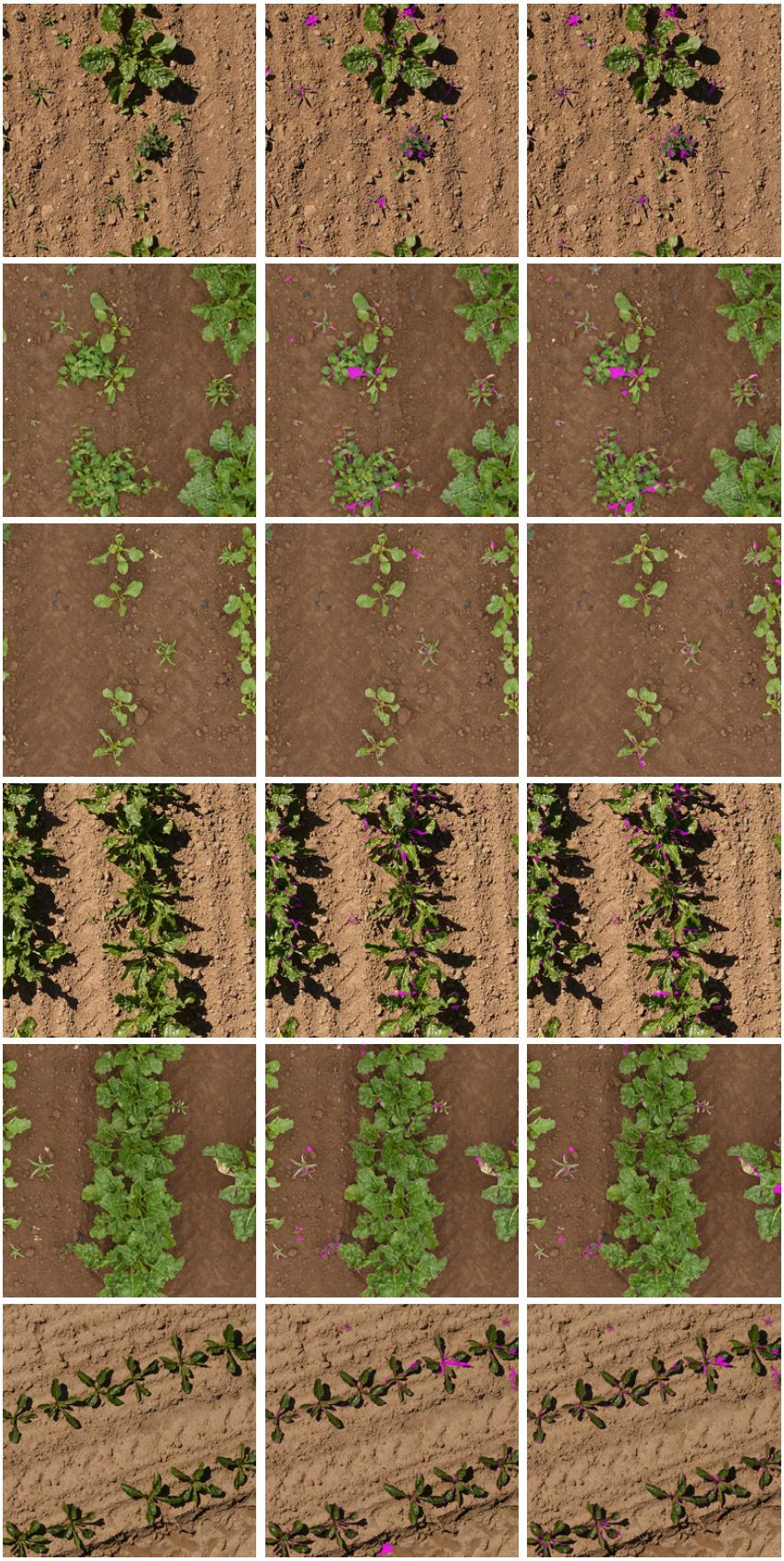
Training setup. First, we employ Panoptic DeepLab [3] with a MobileNetV2 [19] backbone. We initialize the network with weights pre-trained on ImageNet [5] and train the model for 200 epochs with a batch size of 8 using Adam [12]. We employ the WarmupPolyLR scheduler [15] with an initial value of $1 \cdot 10^{-2}$.

Second, we use Mask R-CNN [8] with a ResNet-50 [9] backbone and train the model for 200 epochs with a batch size of 8



(a) Input image (b) Ground truth (c) ERFNet (d) DeepLabV3

Fig. 10. Qualitative results for semantic segmentation on the test set. Colors indicate here the semantic class, where crop pixels are green and weed pixels are red.



(a) Input image (b) ERFNet (c) DeepLabV3

Fig. 11. Qualitative mispredictions for semantic segmentation on the test set. Specifically, we highlight each pixel where the prediction does not match the ground truth in magenta.

Date	Approach	PQ†	PQ$_{\text{crop}}$	PQ$_{\text{weed}}$	IoU$_{\text{soil}}$
05-15-2020	Panoptic DeepLab [3]	59.09	54.81	22.81	99.65
	Mask R-CNN [8]	64.28	66.20	27.12	99.52
	Mask2Former [4]	70.37	73.04	38.52	99.54
05-26-2020	Panoptic DeepLab [3]	59.78	54.66	25.88	98.81
	Mask R-CNN [8]	69.71	76.84	34.67	97.62
	Mask2Former [4]	73.06	80.21	41.92	97.05
06-05-2020	Panoptic DeepLab [3]	58.94	50.52	27.53	98.77
	Mask R-CNN [8]	73.51	79.59	44.83	96.10
	Mask2Former [4]	73.72	78.49	48.38	94.30
2020	Panoptic DeepLab [3]	59.31	54.39	24.16	99.37
	Mask R-CNN [8]	66.72	70.34	31.03	98.78
	Mask2Former [4]	71.49	75.52	40.46	98.49
05-20-2021	Panoptic DeepLab [3]	48.29	34.40	10.94	99.54
	Mask R-CNN [8]	74.44	80.22	43.70	99.41
	Mask2Former [4]	62.22	52.47	34.76	99.44
05-28-2021	Panoptic DeepLab [3]	53.26	43.23	17.25	99.29
	Mask R-CNN [8]	70.29	74.10	37.78	98.99
	Mask2Former [4]	55.32	19.86	46.83	99.27
06-01-2021	Panoptic DeepLab [3]	48.10	39.13	6.51	98.67
	Mask R-CNN [8]	57.64	45.96	29.31	97.66
	Mask2Former [4]	49.72	13.21	37.28	98.68
06-10-2021	Panoptic DeepLab [3]	38.74	17.83	1.48	96.91
	Mask R-CNN [8]	45.21	14.53	30.64	90.45
	Mask2Former [4]	51.84	20.45	40.47	94.59
2021	Panoptic DeepLab [3]	43.73	26.95	6.11	98.12
	Mask R-CNN [8]	55.93	38.81	34.16	94.81
	Mask2Former [4]	54.14	25.82	39.64	96.97

TABLE 11. Baseline results for panoptic segmentation on the test set, separated to 2020, 2021, and each data collection day.

Date	Approach	PQ†	PQ$_{\text{crop}}$	PQ$_{\text{weed}}$	IoU$_{\text{soil}}$
05-15-2020	Panoptic DeepLab [3]	59.91	56.37	23.70	99.66
	Mask R-CNN [8]	65.28	66.61	29.72	99.52
	Mask2Former [4]	70.80	73.21	39.67	99.51
05-26-2020	Panoptic DeepLab [3]	61.63	58.62	27.00	99.27
	Mask R-CNN [8]	68.80	75.21	32.65	98.54
	Mask2Former [4]	73.36	79.93	42.01	98.15
06-05-2020	Panoptic DeepLab [3]	62.37	55.12	33.14	98.86
	Mask R-CNN [8]	76.50	78.78	54.61	96.11
	Mask2Former [4]	79.15	84.61	57.76	95.08
2020	Panoptic DeepLab [3]	61.00	56.54	27.08	99.39
	Mask R-CNN [8]	69.18	71.70	37.30	98.53
	Mask2Former [4]	73.70	77.68	45.22	98.19

TABLE 12. Baseline results for panoptic segmentation on the validation set recorded at different data collection days in 2020.

using the AdamW [14] optimizer. We employ an exponentially decaying schedule [15] with an initial learning rate of $1 \cdot 10^{-4}$.

Finally, we employ Mask2Former [4] based on a ResNet-50 [9] backbone. We initialize the weights based on a model pre-trained on ImageNet [5] and train the model for 200 epochs with a batch size of 8 using the AdamW [14] optimizer. Here we employ the WarmupPolyLR rate scheduler with an initial value of $1 \cdot 10^{-4}$.

Evaluation on the 2020 and 2021 Test Set. Tab. 11 provides quantitative results on the data collected in 2020 and 2021, belonging to the test set. Since the models are only trained on 2020 data, all three approaches suffer a drop in PQ† performance when we test on 2021 data.

Evaluation on the Validation Set. Tab. 12 shows the results of all approaches on the validation set. Compared to the test set results, the gap between Mask R-CNN and Mask2Former is larger, while this difference is on the test set not as pronounced. In particular, the panoptic quality of weeds for Mask R-CNN shows a drop in performance.

Ablation Studies. A key contribution of our paper is a large dataset for semantic image interpretation in the agricultural domain. To demonstrate the necessity of such a large dataset for panoptic segmentation, we train multiple networks with the same hyperparameters but provide each network with a random subset of the training data. Specifically, we create random subsets containing $1\,\%$, $10\,\%$, $25\,\%$, and $50\,\%$ of the original training data and train Mask R-CNN based on each subset.

We report quantitative results based on the test set in Fig. 12. Generally, we observe that a larger training set results in increased performance. This effect is particularly apparent when comparing the quantitative results for small random subsets but less pronounced when increasing the size of the training set. Additionally, we highlight that detecting pixels belonging to soil is less demanding than identifying individual plants belonging to crops and weeds. Thus, only a few training samples can achieve relatively high performance for the former task, while the latter requires substantially more training samples.

Qualitative Results. Fig. 13 shows the qualitative results of the instance segmentation, where different colors indicate different instances. We see a difference between the early and late growth stages. While early growth stages usually show separated plants, later growth stages can be characterized by a larger overlap between plants. At the early growth stage, plant instance masks are generally correctly identified. However, this changes dramatically in later growth stages, where the predicted instance masks between overlapping plants are not well separated. In particular, Mask R-CNN shows very blob-like segmentation that often does not even cover the whole plant, which seems to be caused by the instance prediction branch that predicts masks at a lower resolution and the upsampling to a higher resolution image. We leave the investigation of approaches for refining the instance masks, such as PointRend [13], as an avenue for future work.

The segment boundaries of bottom-up approaches, *i.e.*, Panoptic Deeplab and Mask2Former, are much sharper and follow the plant's shape much better than the results of Mask R-CNN since these approaches predict pixel-wise instance masks. However, both approaches still struggle with the separation of different plants. In particular, the last row of the qualitative results in Fig. 13 shows the limit of commonly employed panoptic segmentation approaches, where all approaches miss a complete plant at the bottom of the image and leaves are wrongly assigned to different crops. As discussed in the paper, we see here an interesting

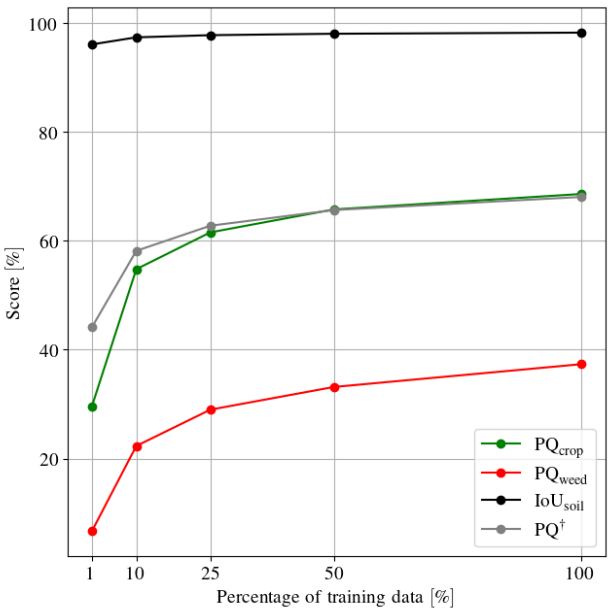


Fig. 12. Quantitative results for panoptic segmentation based on Mask R-CNN models trained on random subsets containing 1 %, 10 %, 25 %, and 50 % of the original training data.

research direction to integrate plant structure in the prediction of instance masks, as sugar beets have a stem location from which the leaves originate.

We further emphasize the issues of each baseline in Fig. 14, where we illustrate each instance as true positive, false positive, or false negative. Particularly, we define any predicted instance that has an IoU greater than 0.5 with any ground truth instance that is not already assigned to a prediction as true positive and visualize it in blue. Oppositely, if a prediction cannot be assigned to any ground truth instance, we consider it as a false positive and illustrate it in a pinkish color. Lastly, we define any ground truth instance that is not assigned to any prediction as a false negative and indicate it in a shade of cyan.

6.3 Detection

For the detection of plants or leaves, we employ Faster R-CNN [16], Mask R-CNN [8], and YOLOv7 [20] as common methods for this task and provide more details about the training procedure in the following.

Training setup. Since Faster R-CNN [16] and Mask R-CNN [8] follow a similar architecture, we apply the same training procedure for both methods. Specifically, we use ResNet-50 [9] as a backbone and train each model for 200 epochs with a batch size of 12 using Adam [12] for optimization. We set the initial learning rate to $1 \cdot 10^{-4}$ and apply an exponentially decaying schedule [15].

For YOLOv7 [20], we trained the network for 300 epochs with a batch size of 16. Additionally, we used the stochastic gradient descent optimizer with an initial learning rate of $1 \cdot 10^{-1}$ and a momentum of 0.937. Furthermore, we employ the OpenCycle [10] learning rate scheduler with the final learning rate of 0.1. We use the 51 layers deep default YOLOv7 backbone. During training, we apply standard data augmentation such as color augmentation, translation, scaling, flipping, mosaic, and mixup, as proposed by the authors of YOLOv7 [20].

Date	Approach	mAP	mAP_{50}	mAP_{75}	AP	
					Crop	Weed
05-15-2020	Faster R-CNN [16]	38.11	62.70	35.14	61.70	14.52
	Mask R-CNN [8]	37.21	62.38	35.70	61.09	13.33
	YOLOv7 [20]	57.63	82.26	59.08	79.16	36.10
05-26-2020	Faster R-CNN [16]	46.92	71.80	50.37	75.72	18.11
	Mask R-CNN [8]	45.90	71.14	46.88	74.36	17.44
	YOLOv7 [20]	63.25	83.44	64.60	89.67	36.83
06-05-2020	Faster R-CNN [16]	52.41	72.77	60.07	74.22	30.61
	Mask R-CNN [8]	48.31	69.54	56.26	71.05	25.56
	YOLOv7 [20]	66.75	81.86	70.43	86.55	46.94
2020	Faster R-CNN [16]	42.35	66.89	42.60	66.77	17.93
	Mask R-CNN [8]	40.86	66.13	41.05	65.16	16.55
	YOLOv7 [20]	60.37	82.31	62.09	82.76	37.98
05-20-2021	Faster R-CNN [16]	49.00	70.46	49.75	78.38	19.62
	Mask R-CNN [8]	49.90	71.33	51.50	78.76	21.03
	YOLOv7 [20]	66.85	85.83	69.23	91.29	42.41
05-28-2021	Faster R-CNN [16]	56.37	80.68	58.48	87.17	25.58
	Mask R-CNN [8]	56.62	81.30	57.07	84.34	28.90
	YOLOv7 [20]	75.56	92.77	76.12	98.58	52.54
06-01-2021	Faster R-CNN [16]	30.68	60.56	27.78	50.03	11.33
	Mask R-CNN [8]	23.37	48.85	17.03	35.86	10.87
	YOLOv7 [20]	63.98	91.06	62.77	94.77	33.20
06-10-2021	Faster R-CNN [16]	13.96	34.74	7.02	11.05	16.87
	Mask R-CNN [8]	11.70	33.41	5.44	5.81	17.59
	YOLOv7 [20]	54.63	81.15	55.00	78.52	30.74
2021	Faster R-CNN [16]	27.10	51.95	22.33	36.26	17.95
	Mask R-CNN [8]	23.83	48.67	19.00	27.77	19.90
	YOLOv7 [20]	62.94	85.41	65.51	86.22	39.66

TABLE 13. Baseline results for plant detection on the test set, separated to 2020, 2021, and each data collection day.

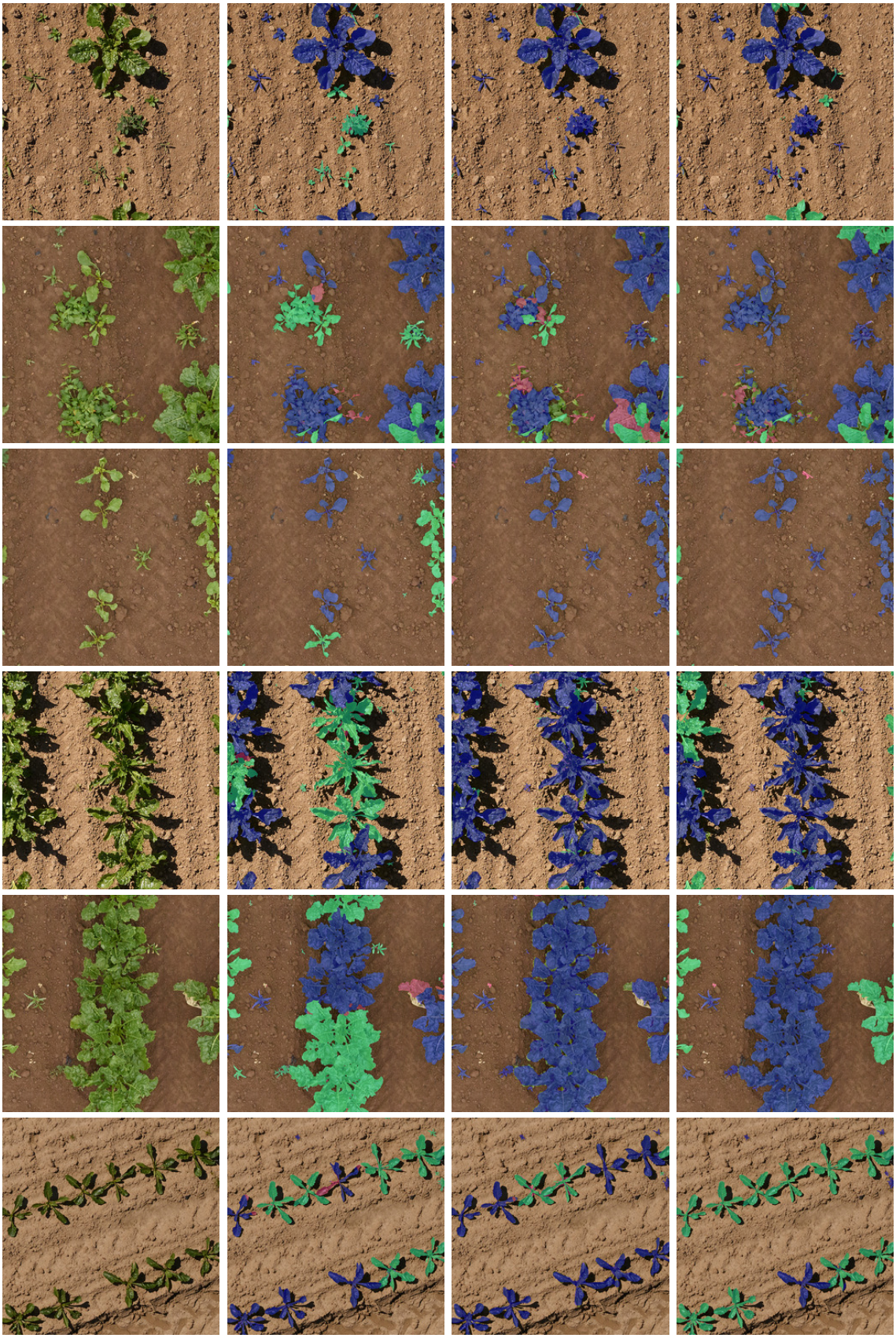
Date	Approach	mAP	mAP_{50}	mAP_{75}	AP	
					Crop	Weed
05-15-2020	Faster R-CNN [16]	37.75	64.08	34.57	59.29	16.21
	Mask R-CNN [8]	36.57	64.23	32.72	57.71	15.43
	YOLOv7 [20]	60.93	86.71	63.21	78.14	43.71
05-26-2020	Faster R-CNN [16]	43.21	64.78	46.54	71.32	15.10
	Mask R-CNN [8]	44.41	65.39	48.15	72.30	16.53
	YOLOv7 [20]	63.48	85.63	66.02	87.07	39.89
06-05-2020	Faster R-CNN [16]	57.63	80.03	64.52	74.42	40.84
	Mask R-CNN [8]	54.93	78.43	60.11	71.87	37.98
	YOLOv7 [20]	76.17	90.61	82.47	88.67	63.66
2020	Faster R-CNN [16]	44.77	69.59	45.92	65.89	23.64
	Mask R-CNN [8]	44.41	70.00	45.03	64.82	24.01
	YOLOv7 [20]	66.95	88.40	71.00	82.90	51.01

TABLE 14. Baseline results for plant detection on the validation set recorded at different data collection days in 2020.

Evaluation on the 2020 and 2021 Test Set. Tab. 13 and Tab. 15 provides quantitative results on the data collected in 2020 and 2021, belonging to the test set for the plant detection and leaf detection tasks. Since the models are only trained on 2020 data, both Faster R-CNN and Mask R-CNN suffer a drop in mAP performance when we test on 2021 data for both the plant and leaf detection task. However, the YOLOv7 approach showed an improvement in performance across all performance metrics.

(a) Input image (b) Ground truth (c) Panoptic DeepLab (d) Mask R-CNN (e) Mask2Former

Fig. 13. Qualitative results for panoptic segmentation on the test set. Colors indicate here different instances and we do not show semantics as these are generally consistent to the semantic segmentation results.



(a) Input image (b) Panoptic DeepLab (c) Mask R-CNN (d) Mask2Former

Fig. 14. Qualitative results distinguishing between correct predictions and mispredictions for panoptic segmentation on the test set. Specifically, we highlight each true positive plant instance in blue, each false positive in a pinkish color, and each false negative in shades of cyan.

Date	Approach	mAP	mAP_{50}	mAP_{75}
05-15-2020	Faster R-CNN [16]	34.95	69.10	31.39
	Mask R-CNN [8]	36.15	69.55	33.48
	YOLOv7 [20]	58.63	90.51	63.65
05-26-2020	Faster R-CNN [16]	32.67	61.86	31.44
	Mask R-CNN [8]	33.62	62.57	32.72
	YOLOv7 [20]	56.91	84.03	61.80
06-05-2020	Faster R-CNN [16]	33.12	61.09	32.89
	Mask R-CNN [8]	34.29	63.08	33.74
	YOLOv7 [20]	59.16	84.76	63.82
2020	Faster R-CNN [16]	33.66	64.54	31.61
	Mask R-CNN [8]	34.60	65.94	33.08
	YOLOv7 [20]	57.90	87.00	62.75
05-20-2021	Faster R-CNN [16]	32.87	67.48	27.83
	Mask R-CNN [8]	33.64	65.62	29.81
	YOLOv7 [20]	61.04	91.53	69.25
05-28-2021	Faster R-CNN [16]	45.00	71.71	52.60
	Mask R-CNN [8]	44.28	70.24	49.14
	YOLOv7 [20]	66.69	95.63	76.40
06-01-2021	Faster R-CNN [16]	38.71	69.64	37.89
	Mask R-CNN [8]	37.86	68.96	36.06
	YOLOv7 [20]	63.56	88.19	72.36
06-10-2021	Faster R-CNN [16]	33.16	64.85	29.90
	Mask R-CNN [8]	32.56	64.62	28.24
	YOLOv7 [20]	54.62	80.97	60.65
2021	Faster R-CNN [16]	34.57	65.94	32.00
	Mask R-CNN [8]	34.34	65.77	30.90
	YOLOv7 [20]	57.78	84.98	65.26

TABLE 15. Baseline results for leaf detection on the test set, separated to 2020, 2021, and each data collection day.

Date	Approach	mAP	mAP_{50}	mAP_{75}
05-15-2020	Faster R-CNN [16]	33.53	66.72	30.62
	Mask R-CNN [8]	34.77	67.68	31.70
	YOLOv7 [20]	58.17	89.73	62.11
05-26-2020	Faster R-CNN [16]	35.06	68.06	31.39
	Mask R-CNN [8]	34.87	68.17	31.63
	YOLOv7 [20]	59.70	89.79	65.06
06-05-2020	Faster R-CNN [16]	37.04	66.47	37.74
	Mask R-CNN [8]	37.99	66.37	39.67
	YOLOv7 [20]	62.41	85.78	68.94
2020	Faster R-CNN [16]	35.40	66.84	33.85
	Mask R-CNN [8]	36.01	67.12	34.75
	YOLOv7 [20]	60.04	88.12	65.08

TABLE 16. Baseline results for leaf detection on the validation set recorded at different data collection days in 2020.

Evaluation on the Validation Set. Tab. 14 provides the validation results for the plant detection, and Tab. 16 provides the results for the leaf detection task on the validation set, which are generally well aligned with the results reported on the test set.

Qualitative Results. In Fig. 16, we show the detection results for plants, where we encode bounding boxes associated with crops in green and associated with weeds in red. In particular, the first row of the qualitative results shows that many approaches struggle to detect small weeds. This is in line with the additional qualitative results in Fig. 17, where we highlight true positive, false positive, and false negative bounding boxes in different colors. Similar to the qualitative results for panoptic segmentation, we specify any predicted bounding box that has an IoU greater than 0.5 with any ground truth bounding box which is not already assigned to a prediction as a true positive. Oppositely, in case the predicted bounding box cannot be assigned to any ground truth bounding box, we consider it as a false positive. Finally, we specify any ground truth bounding box that is not assigned to any prediction as a false negative.

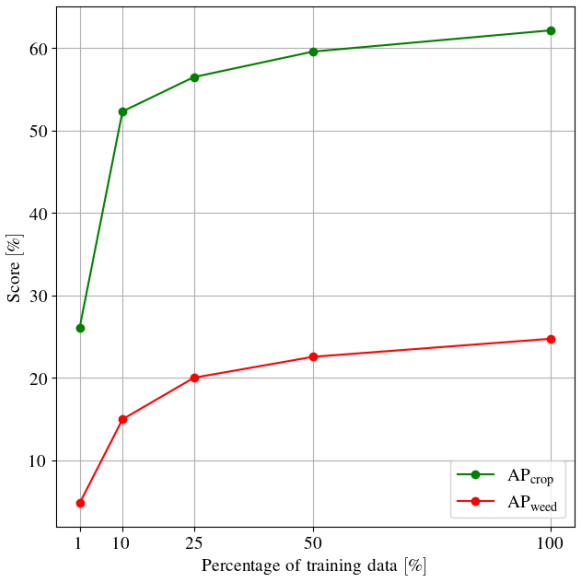


Fig. 15. Quantitative results for plant detection based on Mask R-CNN models trained on random subsets containing 1 %, 10 %, 25 %, and 50 % of the original training data.

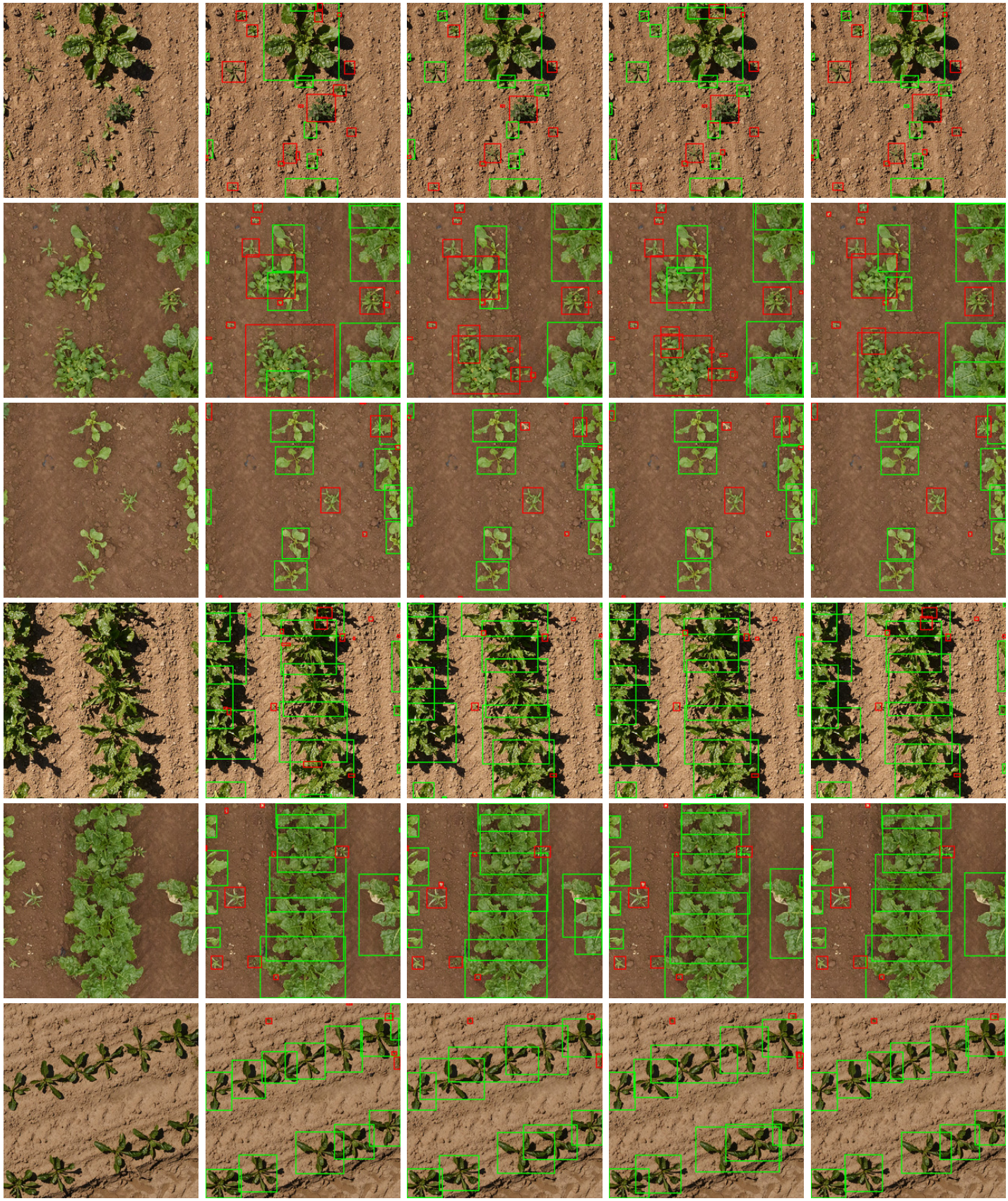
Moreover, we show in in Fig. 18 the detection results for crop leaves, where different colors correspond to the bounding box associated with a specific instance. Like panoptic segmentation, the baselines achieve impressive performance at early growth stages, with moderate overlap between individual leaves. In contrast, the predictions are less accurate for later growth stages, where individual leaves overlap substantially. Here, some leaves are missing, or multiple leaves are detected as a single leaf. This also becomes evident in Fig. 19, where we specify each bounding box as a true positive, false positive, or false negative, emphasized by different colors.

Ablation Studies. In this section, we evaluate the influence of varying sizes of the training set on the performance of plant detection. We investigate multiple Mask R-CNN models trained on reduced sets of training data by creating random subsets containing 1 %, 10 %, 25 %, and 50 % of the original training data. In Fig. 15, we show quantitative results for each subset, emphasizing that more training data increases AP scores for crops and weeds. However, this effect is specifically pronounced in the low data regime.

6.4 Leaf Instance Segmentation

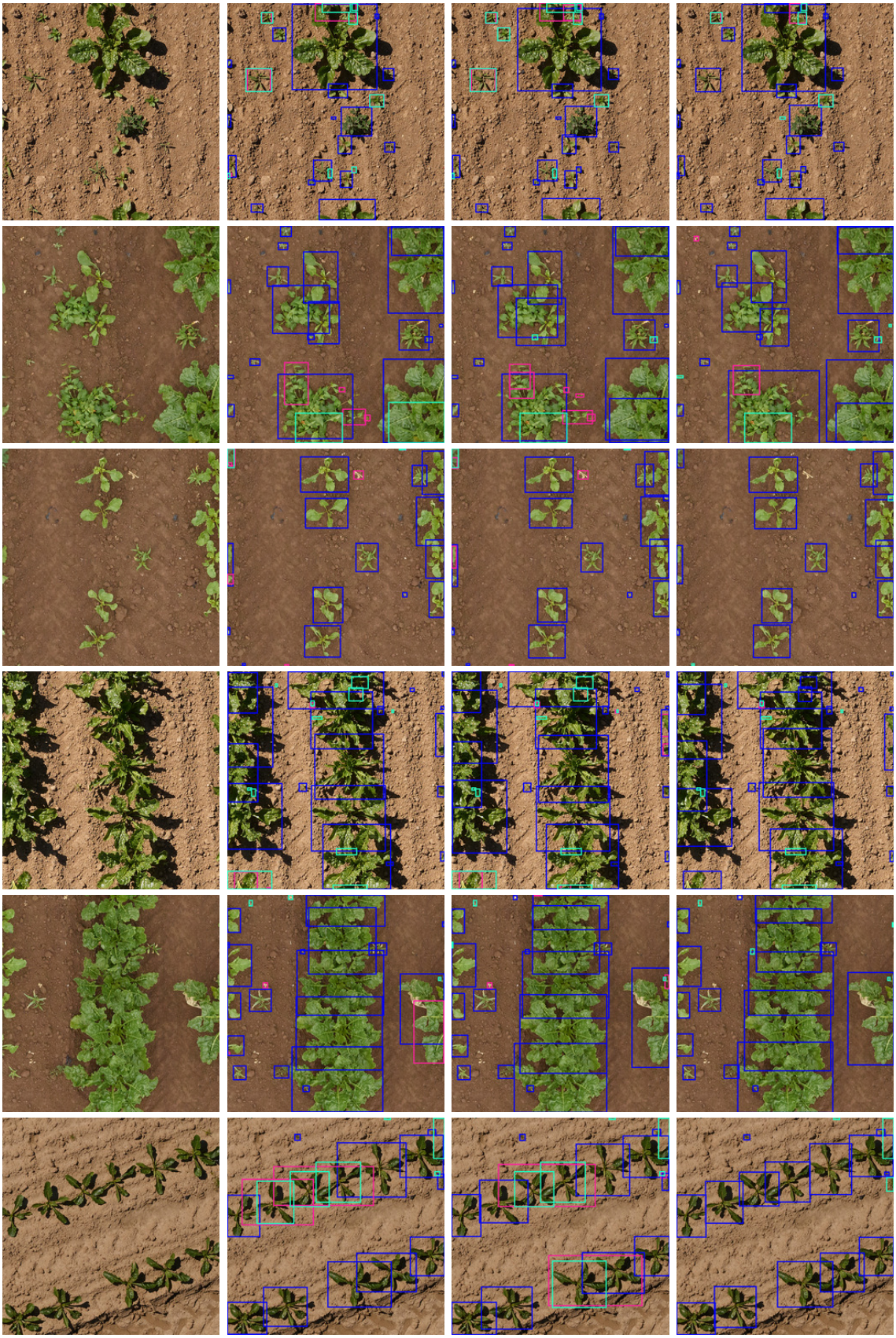
Similarly to panoptic segmentation, we employ Mask R-CNN [8] and Mask2Former [4] for crop leaf instance segmentation and specify the training details in the following.

Training setup. As described for the panoptic segmentation, we train Mask R-CNN [8] with a ResNet-50 [9] backbone for



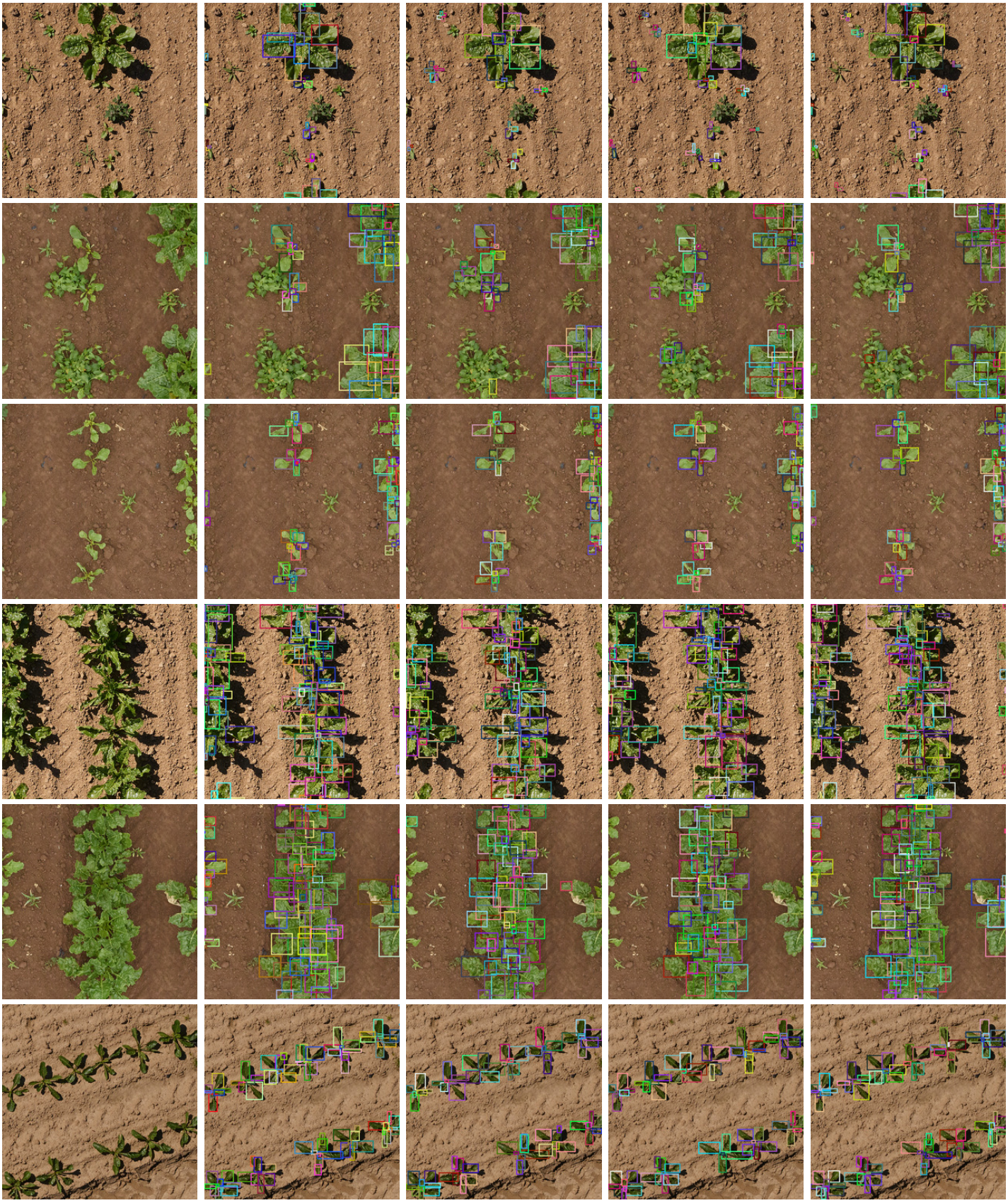
(a) Input image (b) Ground truth (c) Faster R-CNN (d) Mask R-CNN (e) YOLOv7

Fig. 16. Qualitative results for plant detection on the test set, where green bounding boxes correspond to crop detections and red bounding boxes correspond to weed detections.



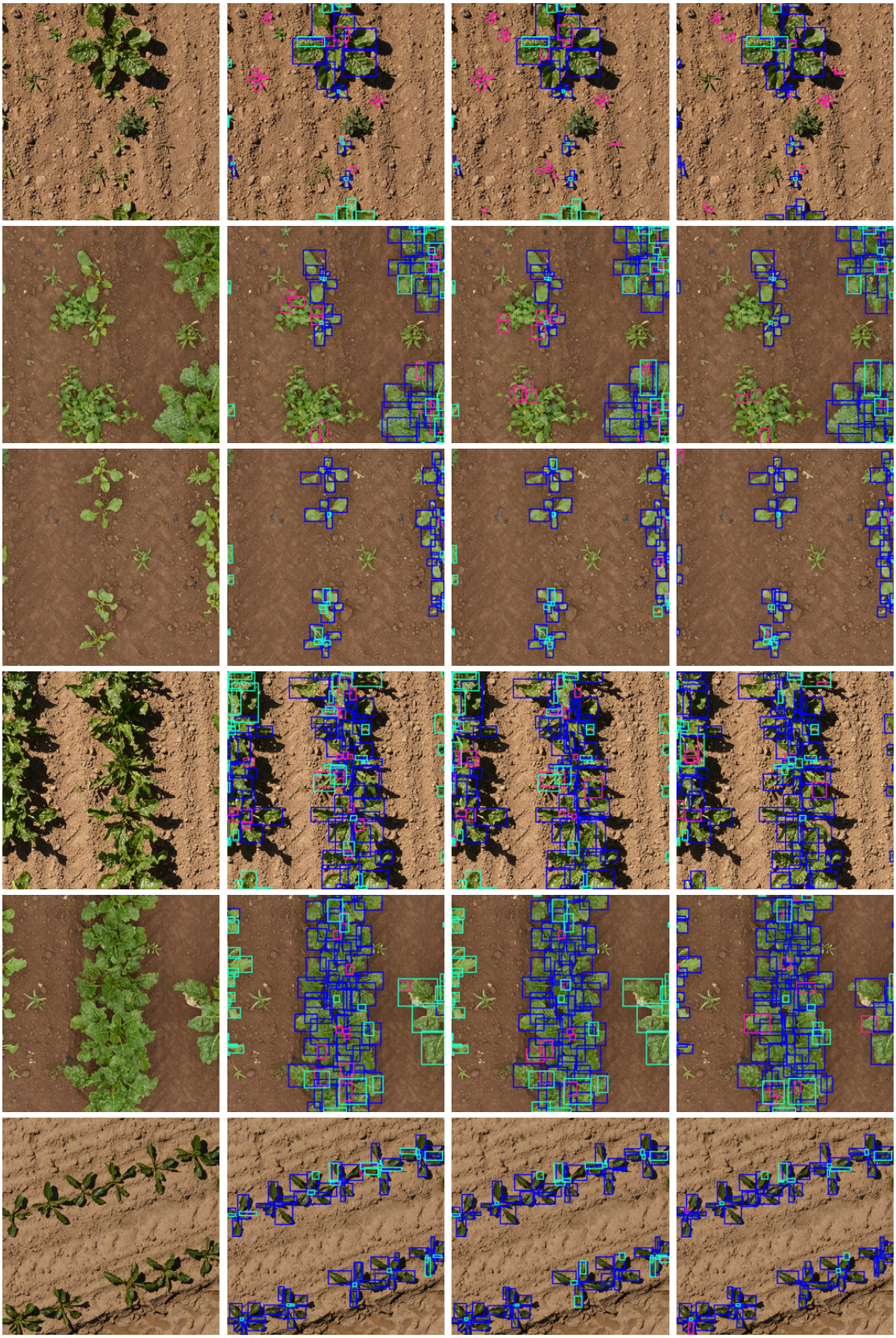
(a) Input image (b) Faster R-CNN (c) Mask R-CNN (d) YOLOv7

Fig. 17. Qualitative results distinguishing between correct predictions and mispredictions for plant detection on the test set. We illustrate each true positive bounding box in blue, each false positive in a pinkish color, and each false negative in shades of cyan.



(a) Input image (b) Ground truth (c) Faster R-CNN (d) Mask R-CNN (e) YOLOv7

Fig. 18. Qualitative results for leaf detection on the test set. Here different colors indicate different instances.



(a) Input image (b) Faster R-CNN (c) Mask R-CNN (d) YOLOv7

Fig. 19. Qualitative results distinguishing between correct predictions and mispredictions for leaf detection on the test set. We illustrate each true positive bounding box in blue, each false positive in a pinkish color, and each false negative in shades of cyan.

Date	Approach	PQ_{leaf}
05-15-2020	Mask R-CNN [8]	59.27
	Mask2Former [4]	58.65
05-26-2020	Mask R-CNN [8]	58.84
	Mask2Former [4]	60.01
06-05-2020	Mask R-CNN [8]	61.17
	Mask2Former [4]	60.84
2020	Mask R-CNN [8]	59.32
	Mask2Former [4]	59.23
05-20-2021	Mask R-CNN [8]	63.37
	Mask2Former [4]	44.50
05-28-2021	Mask R-CNN [8]	69.69
	Mask2Former [4]	38.11
06-01-2021	Mask R-CNN [8]	67.75
	Mask2Former [4]	36.74
06-10-2021	Mask R-CNN [8]	62.48
	Mask2Former [4]	38.29
2021	Mask R-CNN [8]	64.20
	Mask2Former [4]	39.30

TABLE 17. Baseline results for leaf instance segmentation on test set, separated to 2020, 2021, and each data collection day.

200 epochs with a batch size of 12 using the AdamW [14] optimizer. We set the initial learning rate to $1 \cdot 10^{-4}$ and employ an exponentially decaying schedule [15].

Similarly, we use Mask2Former [4] with ResNet-50 [1], [9] as a backbone and optimize the weights of the model with AdamW [14]. Additionally, we employ the WarmupPolyLR rate scheduler[15] with an initial value of $1 \cdot 10^{-4}$, use weights pre-trained on ImageNet and train for 200 epochs with a batch size of 8 images.

Evaluation on the 2020 and 2021 Test Set. Tab. 17 provides quantitative results on the data collected in 2020 and 2021, belonging to the test set. Since the we train the models only on 2020 data, Mask2Former suffer a drop in mAP performance when we test on 2021 data for the leaf instance segmentation task. However, the Mask R-CNN approach shows an improvement in performance.

Evaluation on the Validation Set. In Tab. 18, we present the performance of each baseline on the validation set. Both methods achieve a slightly increased performance on the validation set compared to the test set. However, the results are consistent, *i.e.*, Mask2Former [4] still achieves a better performance than Mask R-CNN [8].

Qualitative Results. In Fig. 20, we show the qualitative results of the leaf instance segmentation for crops, where different colors indicate different instances. Specifically, the instance masks of Mask R-CNN [8] have a blob-like shape that is less accurate compared with Mask2Former [4]. We attribute this to the instance segmentation head of Mask R-CNN, which predicts a low resolution mask that is upsampled for the full resolution image. Again, we leave investigation of refinement strategies [13] for future work. We support these results in Fig. 21, where we highlight true positive, false positive, and false negative leaf instances in different colors.

Date	Approach	PQ_{leaf}
05-15-2020	Mask R-CNN [8]	59.78
	Mask2Former [4]	60.25
05-26-2020	Mask R-CNN [8]	61.20
	Mask2Former [4]	63.58
06-05-2020	Mask R-CNN [8]	65.14
	Mask2Former [4]	65.30
2020	Mask R-CNN [8]	61.50
	Mask2Former [4]	62.31

TABLE 18. Baseline results for leaf instance segmentation on the validation set recorded at different data collection days in 2020.

Date	Approach	$PQ^{\dagger}$	PQ	PQ_{crop}	PQ_{leaf}	IoU	
						Weed	Soil
05-15-2020	HAPT [17]	64.83	54.72	62.79	46.64	50.36	99.53
	Weyler [21]	-	40.64	38.88	42.41	-	-
05-26-2020	HAPT [17]	63.36	47.78	48.92	46.65	60.74	97.14
	Weyler [21]	-	44.38	43.52	45.24	-	-
06-05-2020	HAPT [17]	63.70	43.41	41.39	45.43	73.14	94.85
	Weyler [21]	-	32.73	32.09	33.37	-	-
2020	HAPT [17]	66.50	51.79	57.04	46.54	63.89	98.55
	Weyler [21]	-	40.99	39.57	42.40	-	-
05-20-2021	HAPT [17]	58.15	47.03	56.99	37.06	39.06	99.47
	Weyler [21]	-	44.33	39.70	48.95	-	-
05-28-2021	HAPT [17]	66.48	52.71	46.70	58.72	61.22	99.27
	Weyler [21]	-	55.86	50.76	60.97	-	-
06-01-2021	HAPT [17]	55.07	42.39	28.54	56.24	36.96	98.54
	Weyler [21]	-	41.38	29.79	52.96	-	-
06-10-2021	HAPT [17]	50.96	33.21	15.10	51.33	41.11	96.32
	Weyler [21]	-	26.98	16.19	37.77	-	-
2021	HAPT [17]	54.52	39.48	28.96	49.99	41.34	97.80
	Weyler [21]	-	35.74	26.74	44.74	-	-

TABLE 19. Baseline results for hierarchical panoptic segmentation on the test set, separated to 2020, 2021, and each data collection day.

6.5 Hierarchical Panoptic Segmentation

We select HAPT [17] and the method proposed by Weyler *et al.* [21] as baselines for hierarchical panoptic segmentation, where both models use ERFNet [18] as a backbone. Next, we provide more details about the training procedures for both methods.

Training setup. We train HAPT for 200 epochs with a batch size of 16 and employ AdamW [14] during optimization. Additionally, we set the step learning rate scheduler with an initial value of $4 \cdot 10^{-4}$ for the backbone, and three exponential schedulers with initial learning rates of $\left(4 \cdot 10^{-4}, 8 \cdot 10^{-4}, 8 \cdot 10^{-4}\right)$ for the semantic, plant instance, and leaf instance decoders, respectively. During training, we apply dropout with a probability of 0.15

Furthermore, we train the model by Weyler *et al.* [21] for 512 epochs with a batch size of 1 and use Adam [12] for optimization. We set the initial learning rate to $1 \cdot 10^{-}$ and subsequently apply a polynomial learning rate decay $\left(1 - \frac{e}{512}\right)^{0.9}$, where e is the current epoch.

For both methods, we do not apply any data augmentations and keep the original image size of the dataset.

Evaluation on the 2020 and 2021 Test Set. Tab. 19 provides quantitative results on the data collected in 2020 and 2021,

Date	Approach	PQ†	PQ	PQ_{crop}	PQ_{leaf}	IoU	
						Weed	Soil
05-15-2020	HAPT [17]	65.14	56.08	64.41	47.75	48.89	99.49
	Weyler [21]	-	44.02	43.07	44.96	-	-
05-26-2020	HAPT [17]	65.33	52.65	55.70	49.60	58.15	97.87
	Weyler [21]	-	48.20	47.86	48.53	-	-
06-05-2020	HAPT [17]	63.32	45.83	43.17	48.49	67.09	94.53
	Weyler [21]	-	31.93	29.28	34.57	-	-
2020	HAPT [17]	66.60	52.63	56.91	48.35	63.14	97.99
	Weyler [21]	-	41.76	40.50	43.02	-	-

TABLE 20. Baseline results for hierarchical panoptic segmentation on the validation set recorded at different data collection days in 2020.

belonging to the test set. The HAPT approaches suffer a drop in PQ† and PQ performance when we test on 2021 data for the leaf instance segmentation task. Similiarly, the approach by Weyler *et al*. [21] also suffers a drop in PQ performance when tested on 2021 data. The drop in performance for 2021 data in both approaches is expected since we train both approaches only on 2020 data.

Evaluation on the Validation Set. In consensus with the results on the test set, we observe that HAPT [17] shows superior performance compared with the method Weyler *et al*. [21], see Tab. 20. However, the performance of HAPT is slightly worse compared to its results on the test, while the other method by Weyler *et al*. achieves a minor performance increase on the validation set.

Qualitative Results. Since the methods for hierarchical panoptic segmentation perform a simultaneous instance segmentation of crop leaf and plant instances, we present the qualitative results for each separately in Fig. 22 and Fig. 24, respectively. We observe for both baselines that the quality of predicted instance masks for crops leaves and plants generally decreases with increasing growth stages. Specifically, the baselines struggle to distinguish instances with substantial overlap. As mentioned in the paper, we see a need for models that target these challenging scenarios, *e.g*., by incorporating the plant structure more explicitly. We consider this challenging since the number of leaves per plant varies highly. Thus, plants do not have a strong prior assumption about their total number of leaves, contrary to human pose estimation, where the number of parts per instance is often a constant.

We support these results in Fig. 23 and Fig. 25, where we show true positive, false positive, and false negative instances of plants and leaves, respectively. Here, we observe many false negative instances at late growth stages, i.e., ground instances that cannot be assigned to any prediction, illustrated in shades of cyan.

7 Extended Dataset Statistics

As the recorded field belongs to a farm of the University of Bonn, we can conduct field studies and study perception systems under varying conditions with respect to the application of herbicides, resulting in different scenarios with fully (conventional), partial (80% herbicides), and non-herbicide field conditions. We annotate images of the field regions treated with conventional and reduced herbicidal weed control. We emphasize that the distinct application of agrochemicals has a substantial effect on the number of weeds present in the field. To this end, we report in Tab. 21 the total number of weeds in the training, validation, and testing split for each data collection day and each field treatment separately.

Furthermore, we provide in Tab. 22 more details about the annotated images collected in different years at various dates and split into three distinct sets, i.e., train, val, and test.. Specifically, we report the average ratio between all pixels belonging to the foreground, i.e., crops and weeds, and the background. Additionally, we compute the average canopy cover of all plants and, ultimately, the average number of leaves per crop.

Split	Date	Partially-Herbicided	Fully-Herbicided
Train	05-15-2020	2259	603
	05-26-2020	2410	–
	06-05-2020	2869	–
Validation	05-15-2020	658	591
	05-26-2020	1097	–
	06-05-2020	1580	–
Test	05-15-2020	1104	559
	05-26-2020	1522	–
	06-05-2020	623	–
	05-20-2021	694	–
	05-28-2021	138	–
	06-01-2021	100	–
	06-10-2021	434	–
Total		15488	1753

TABLE 21. Statistics about the number of weeds in the train, val, and test split separated by each data collection day and field treatment.

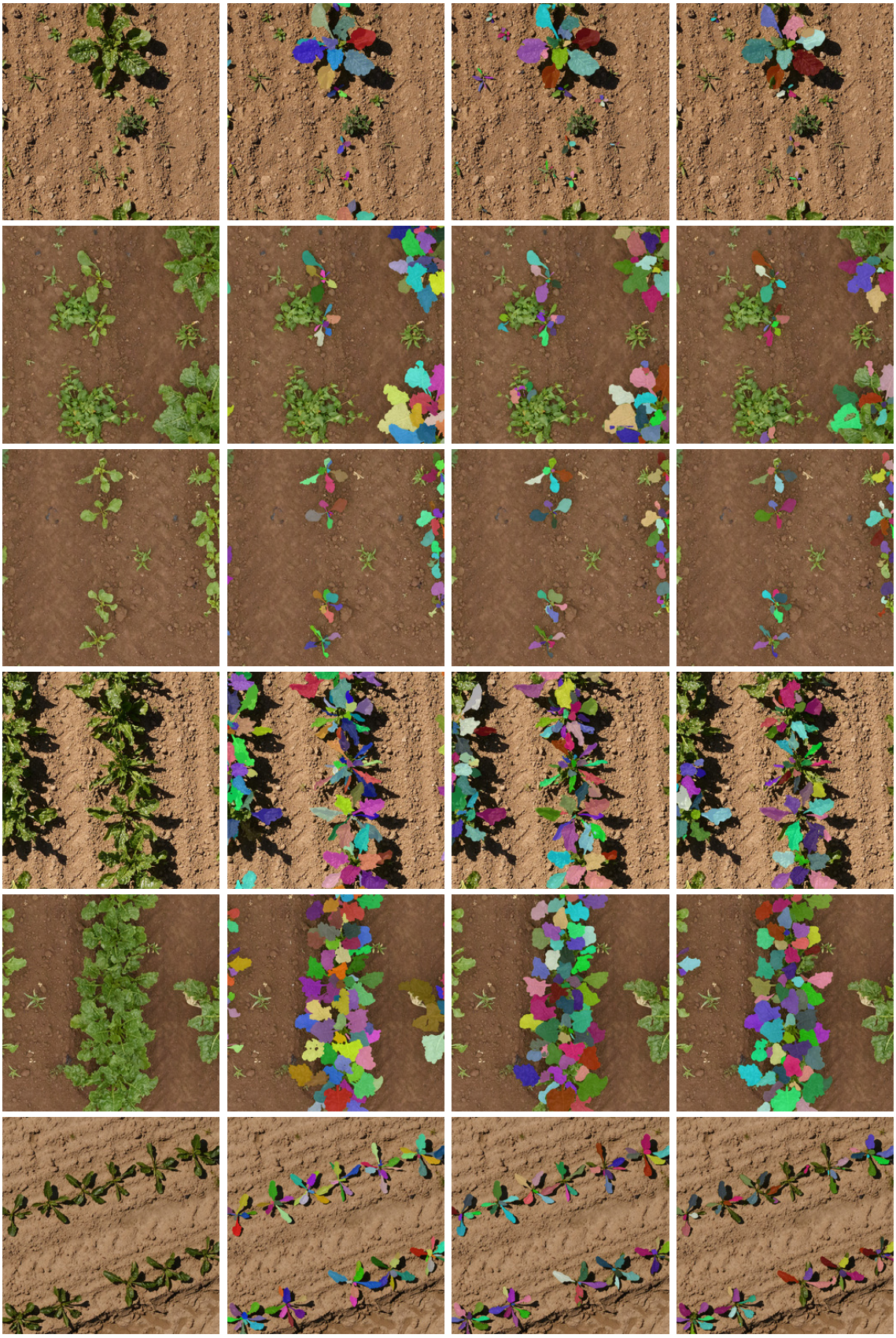
		$^{fg}/_{bg}$	Canopy Cover $[cm^2]$	Number of Leaves
Year	2020	0.11	128.51	6.37
	2021	0.19	141.25	7.11
Date	05-15-2020	0.03	44.63	4.90
	05-26-2020	0.12	137.88	6.95
	06-05-2020	0.23	267.90	8.39
	05-20-2021	0.02	21.42	5.47
	05-28-2021	0.05	46.10	6.56
	06-01-2021	0.09	72.88	6.85
	06-10-2021	0.27	201.73	7.68
Split	train	0.12	147.18	6.70
	val	0.09	119.31	6.17
	test	0.10	104.14	6.07

TABLE 22. Averaged dataset statistics across different years, data collection days, and provided splits.

8 Additional Unlabeled Data

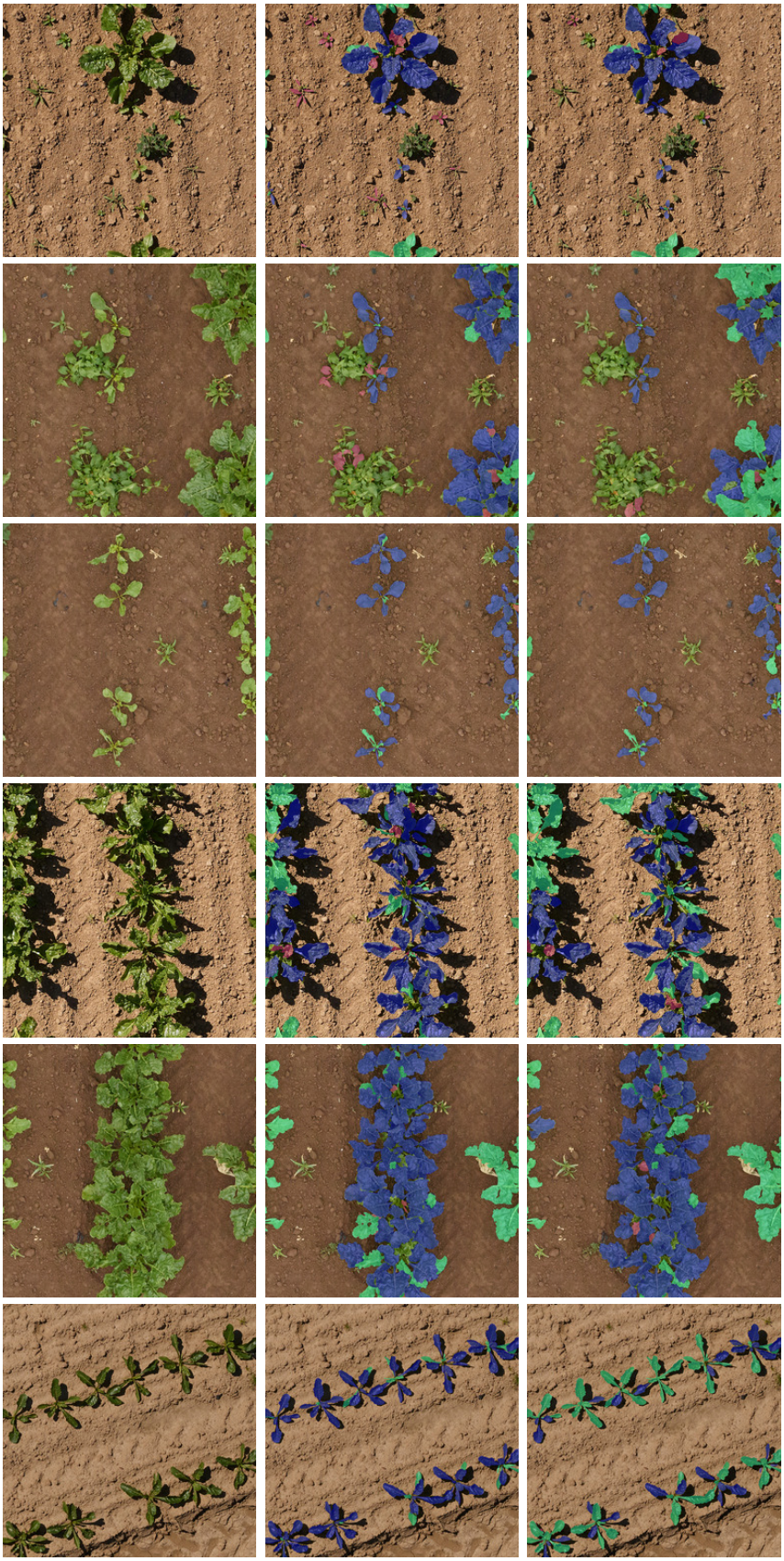
Together with the annotated data and the dedicated splits, we provide also unlabeled data. We hope that this additional data can be exploited using a semi-supervised approach. Furthermore, we think that the recent interest in self-supervised representation learning [2], [7], [6] is an interesting avenue for further research and in this way can be supported through our dataset.

The geographical location of the unlabeled images is the same field plot as the train and validation sets of the labeled images. In addition to the plots used for the labeled data, we include plots without any herbicidal weed control for the unlabeled images. We included either "train" or "val" in the provided unlabeled image filenames to differentiate which plots the images are of. Specifically, we captured these images on seven distinct date in 2020, i.e., April 25, May 3, May 15, May 26, June 5, June 12, July 2. Of these seven days, three days, i.e., 15th of May 2020, 26th of May 2020, and 5th of June, are taken from the same data collection run as the labeled data. However, we also include the date each image was taken in their respective filenames.



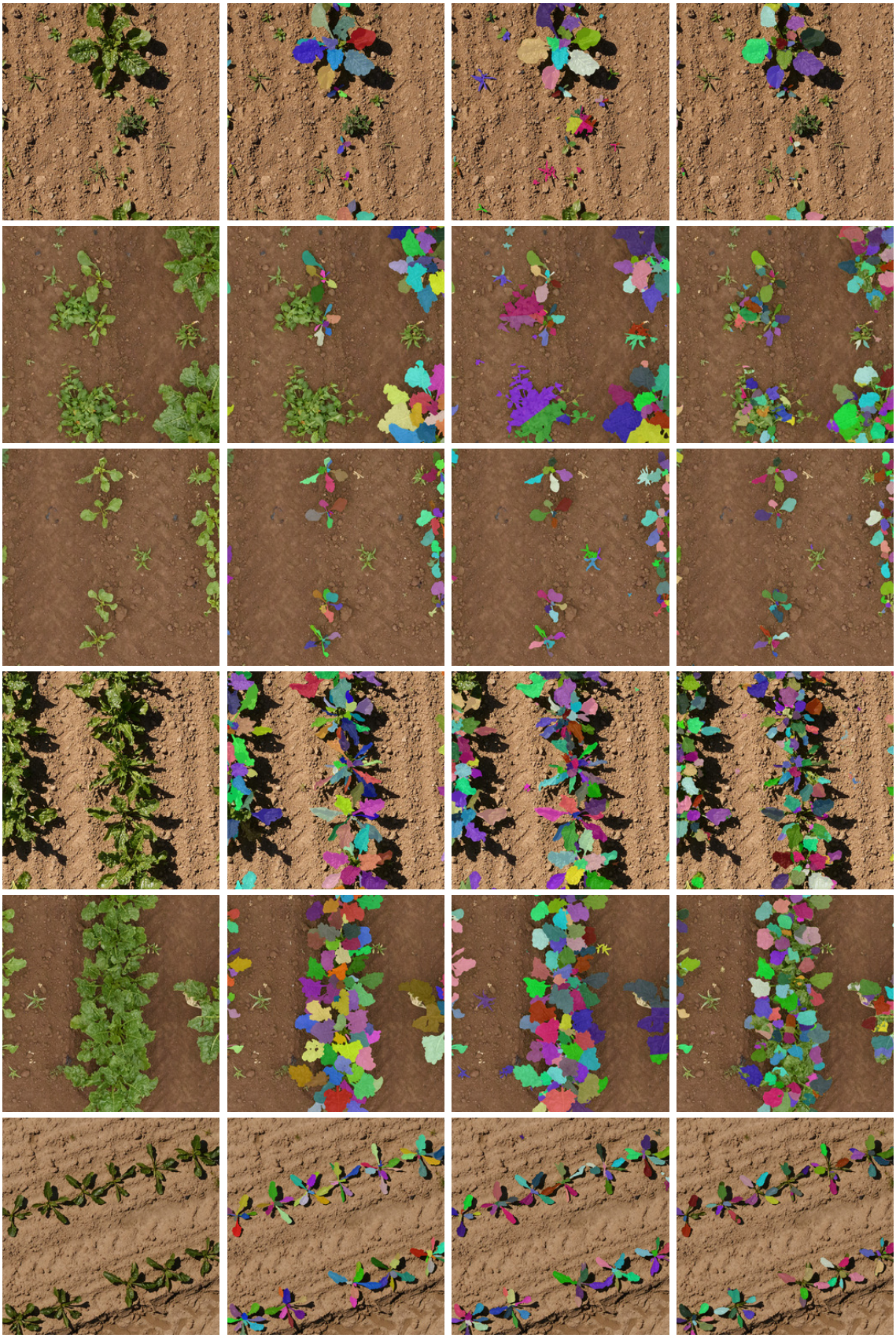
(a) Input image (b) Ground truth (c) Mask R-CNN (d) Mask2Former

Fig. 20. Qualitative results for leaf instance segmentation on the test set.



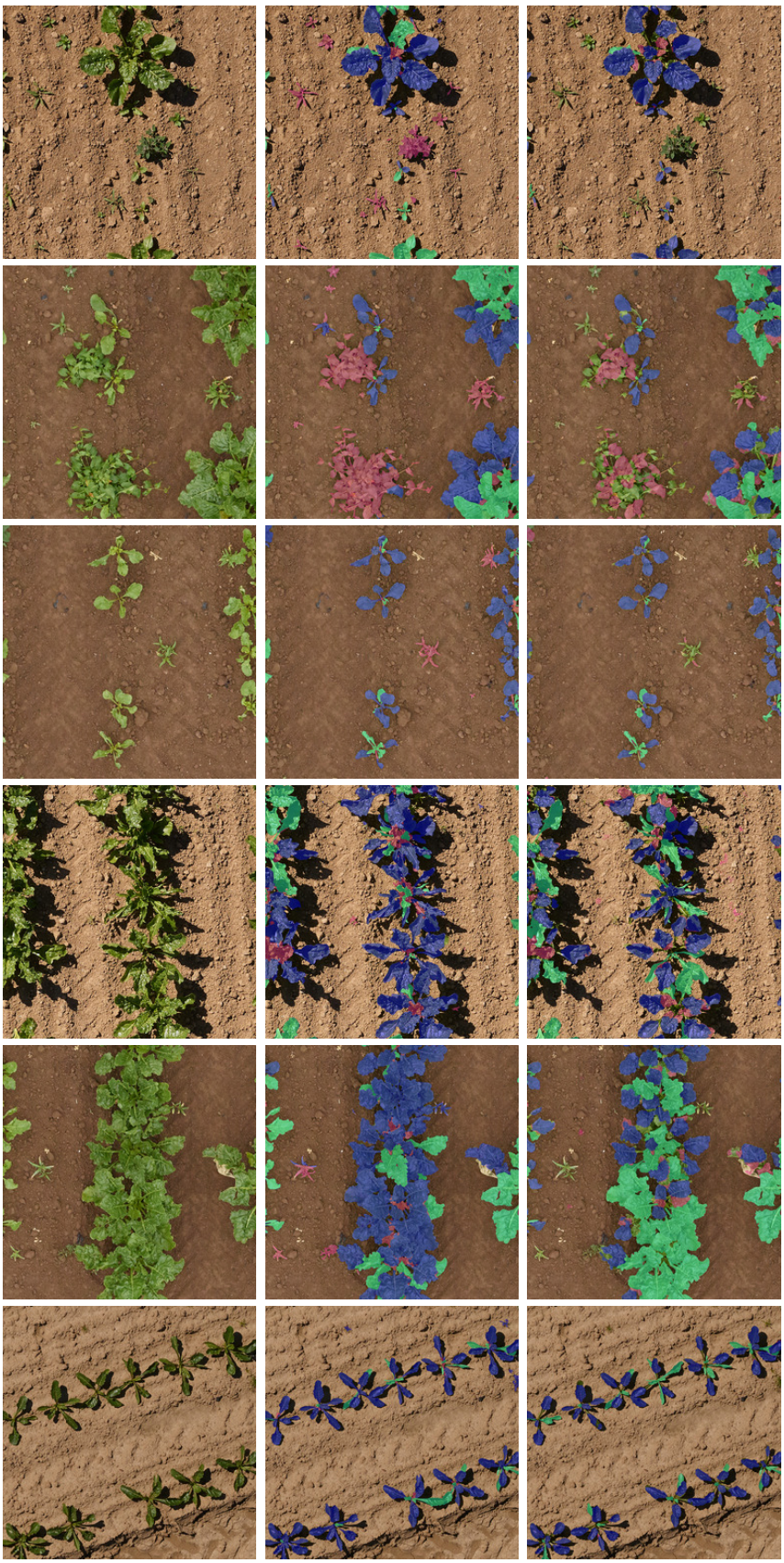
(a) Input image (b) Mask R-CNN (c) Mask2Former

Fig. 21. Qualitative results distinguishing between correct predictions and mispredictions for leaf instance segmentation on the test set. We illustrate each true positive leaf instance in blue, each false positive in a pinkish color, and each false negative in shades of cyan.



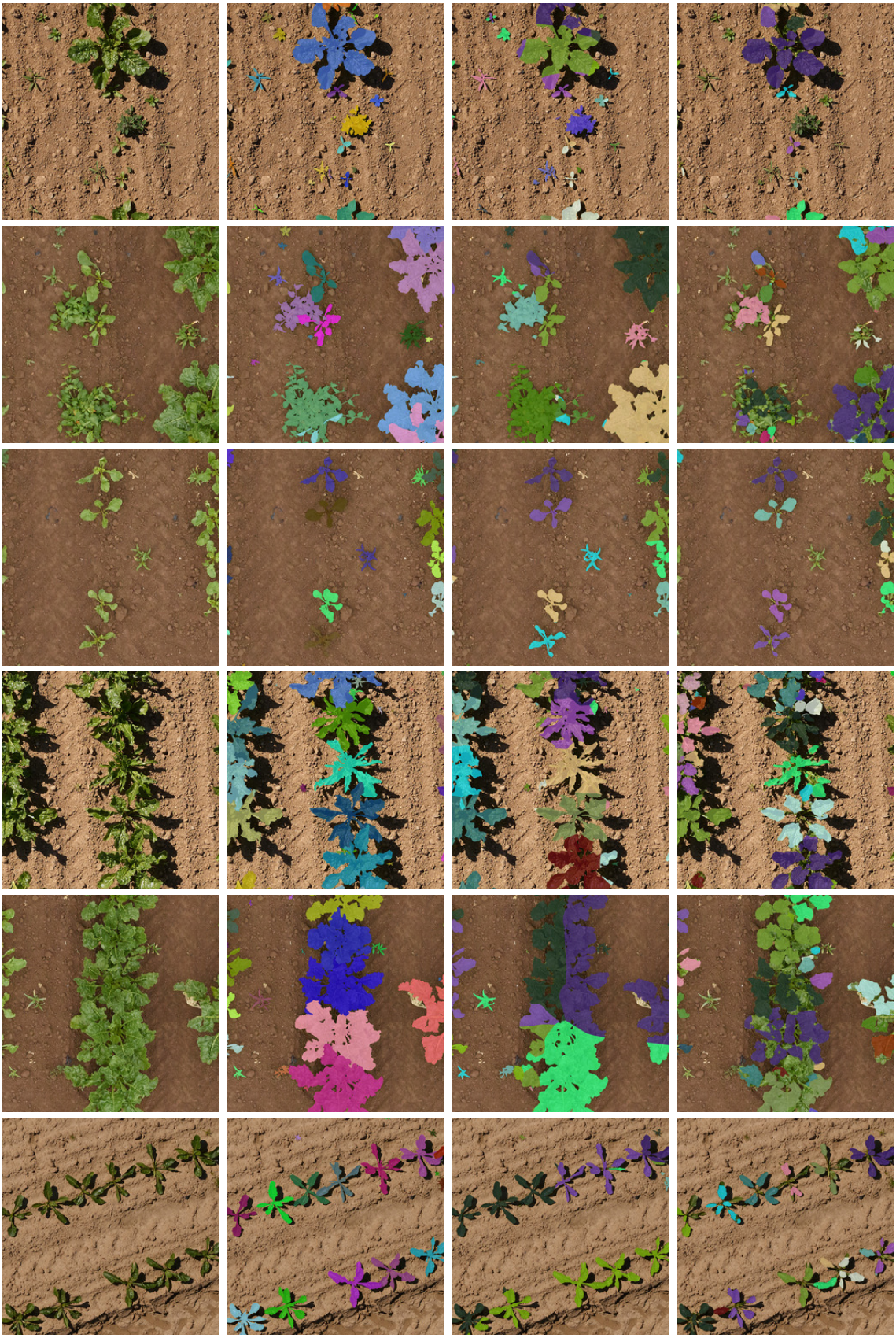
(a) Input image (b) Ground truth (c) HAPT (d) Weyler

Fig. 22. Qualitative results for hierarchical panoptic segmentation targeting crop leaf instances on the test set. Different colors of the masks indicate different instances.



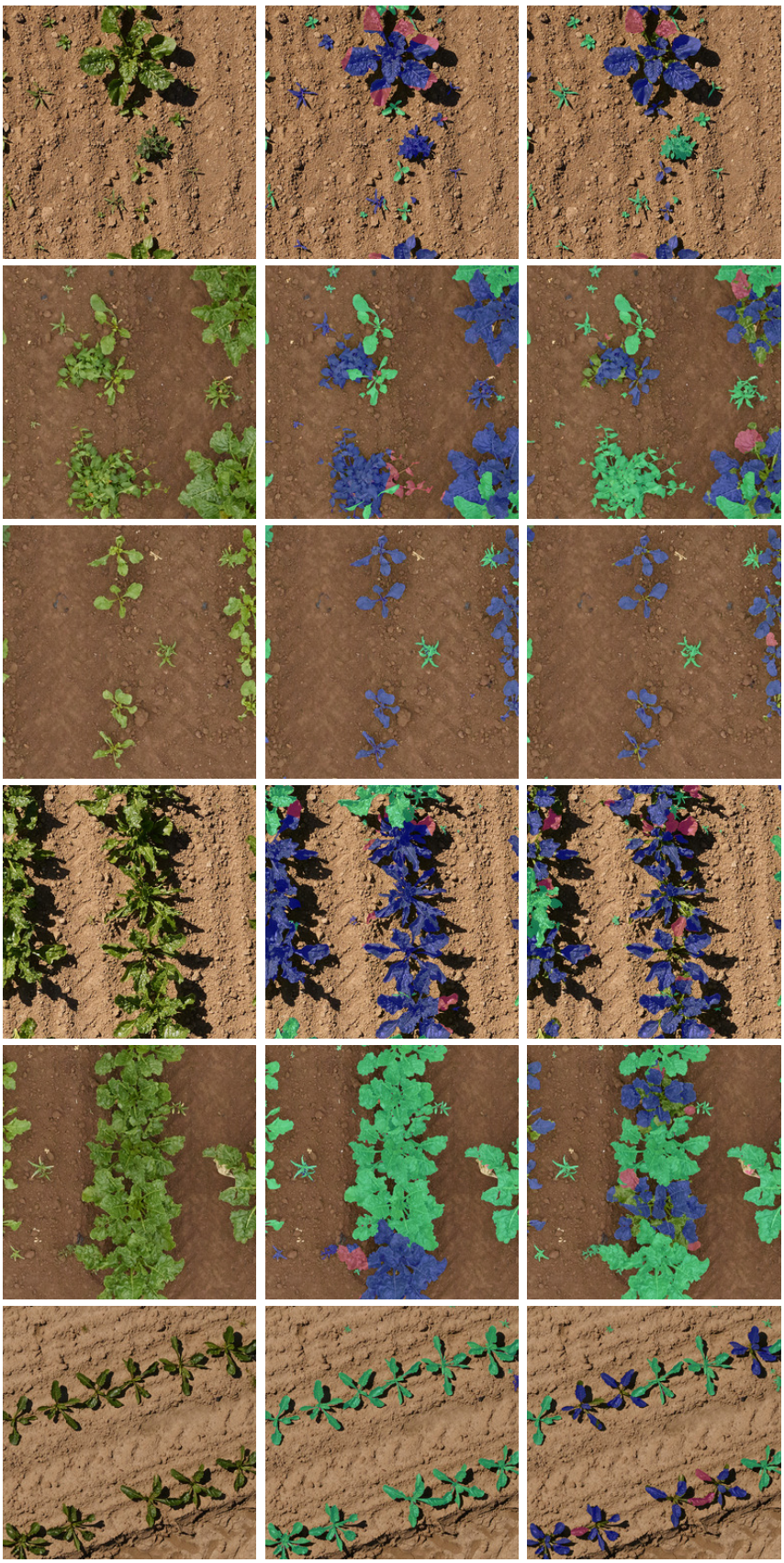
(a) Input image (b) HAPT (c) Weyler

Fig. 23. Qualitative results distinguishing between correct predictions and mispredictions for hierarchical panoptic segmentation targeting crop leaf instances on the test set. We illustrate each true positive crop leaf instance in blue, each false positive in a pinkish color, and each false negative in shades of cyan.



(a) Input image (b) Ground truth (c) HAPT (d) Weyler

Fig. 24. Qualitative results for hierarchical panoptic segmentation targeting crop instances on the test set. We emphasize that HAPT additionally predicts weed instances, which not the case for approach proposed by Weyler.



(a) Input image (b) HAPT (c) Weyler

Fig. 25. Qualitative results distinguishing between correct predictions and mispredictions for hierarchical panoptic segmentation targeting crop instances on the test set. We illustrate each true positive instance in blue, each false positive in a pinkish color, and each false negative in shades of cyan. Since HAPT additionally predicts weed instances, we also consider these plants in the visualization of mispredictions.

REFERENCES

[1] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking Atrous Convolution for Semantic Image Segmentation," *arXiv preprint:1706.05587*, 2017. 3, 4, 16

[2] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," in *Proc. of the Intl. Conf. on Machine Learning (ICML)*, 2020. 17

[3] B. Cheng, M. D. Collins, Y. Zhu, T. Liu, T. S. Huang, H. Adam, and L.-C. Chen, "Panoptic-DeepLab: A Simple, Strong, and Fast Baseline for Bottom-Up Panoptic Segmentation," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 4, 7

[4] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention Mask Transformer for Universal Image Segmentation," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 7, 11, 16

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2009. 4, 7

[6] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum Contrast for Unsupervised Visual Representation Learning," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 17

[7] K. He, R. Girshick, and P. Dollar, "Rethinking ImageNet Pre-training," in *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision (ICCV)*, 2019. 17

[8] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. of the IEEE Intl. Conf. on Computer Vision (ICCV)*, 2017. 4, 7, 8, 11, 16

[9] 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. 4, 7, 8, 11, 16

[10] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of Tricks for Image Classification with Convolutional Neural Networks," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 8

[11] S. Jadon, "A survey of loss functions for semantic segmentation," *arXiv preprint: 2006.14822v4*, 2020. 4

[12] D. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Proc. of the Intl. Conf. on Learning Representations (ICLR)*, 2015. 4, 8, 16

[13] A. Kirillov, Y. Wu, K. He, and R. Girshick, "Pointrend: Image segmentation as rendering," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 7, 16

[14] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. of the Intl. Conf. on Learning Representations (ICLR)*, 2019. 7, 16

[15] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Proc. of the Conf. on Neural Information Processing Systems (NeurIPS)*, 2019, pp. 8026–8037. 1, 4, 7, 8, 16

[16] 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 Conf. on Neural Information Processing Systems (NIPS)*, 2015. 8, 11

[17] G. Roggiolani, M. Sodano, T. Guadagnino, F. Magistri, J. Behley, and C. Stachniss, "Hierarchical Approach for Joint Semantic, Plant Instance, and Leaf Instance Segmentation in the Agricultural Domain," in *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2023. 16, 17

[18] E. Romera, J. M. Alvarez, L. M. Bergasa, and R. Arroyo, "Erfnet: Efficient residual factorized convnet for real-time semantic segmentation," *IEEE Trans. on Intelligent Transportation Systems (ITS)*, vol. 19, no. 1, pp. 263–272, 2017. 3, 4, 16

[19] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018. 4

[20] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," *arXiv preprint: 2207.02696*, 2022. 8, 11

[21] J. Weyler, F. Magistri, P. Seitz, J. Behley, and C. Stachniss, "In-Field Phenotyping Based on Crop Leaf and Plant Instance Segmentation," in *Proc. of the IEEE Winter Conf. on Applications of Computer Vision (WACV)*, 2022. 16, 17