The CropAndWeed Dataset: a Multi-Modal Learning Approach for Efficient Crop and Weed Manipulation

Daniel Steininger Andreas Trondl Gerardus Croonen Julia Simon
Verena Widhalm
AIT Austrian Institute of Technology
{daniel.steininger,andreas.trondl.fl,gerardus.croonen,verena.widhalm,julia.simon}@ait.ac.at

Abstract

Precision Agriculture and especially the application of automated weed intervention represents an increasingly essential research area, as sustainability and efficiency considerations are becoming more and more relevant. While the potentials of Convolutional Neural Networks for detection, classification and segmentation tasks have successfully been demonstrated in other application areas, this relatively new field currently lacks the required quantity and quality of training data for such a highly data-driven approach. Therefore, we propose a novel large-scale image dataset specializing in the fine-grained identification of 74 relevant crop and weed species with a strong emphasis on data variability. We provide annotations of labeled bounding boxes, semantic masks and stem positions for about 112k instances in more than 8k high-resolution images of both real-world agricultural sites and specifically cultivated outdoor plots of rare weed types. Additionally, each sample is enriched with an extensive set of meta-annotations regarding environmental conditions and recording parameters. We furthermore conduct benchmark experiments for multiple learning tasks on different variants of the dataset to demonstrate its versatility and provide examples of useful mapping schemes for tailoring the annotated data to the requirements of specific applications. In the course of the evaluation, we furthermore demonstrate how incorporating multiple species of weeds into the learning process increases the accuracy of crop detection. Overall, the evaluation clearly demonstrates that our dataset represents an essential step towards overcoming the data gap and promoting further research in the area of Precision Agriculture.

1. Introduction

In times of worldwide population growth, the agricultural industry faces a growing demand for food crops, while reducing its impact on the environment and human health is simultaneously gaining priority. One of the tasks offering the highest potential is the efficient removal of weeds, which otherwise compete with cultivated crops for resources such as sunlight, water, space and nutrients, thereby significantly reducing their overall yield [25, 9]. Instead of the commonly used technique of applying herbicides at a large scale to the entire cultivated area, Precision Agriculture, especially in the form of automatic localization and classification of crops and weeds presents the opportunity to significantly increase both the efficiency and sustainability of the process by precisely applying herbicides to specific plants or even removing weeds mechanically without the use of chemicals. Furthermore, similar methods can be used for other tasks such as growth tracking, automatic harvesting and even recognizing plant pests or diseases [34, 32].

Figure 1. Representative annotation example with semantic masks (top right), bounding boxes (bottom left) and stem positions (bottom right) depicting 4 of the 74 crop and weed species.

Both Convolutional Neural Networks (CNNs) [14, 11, 21] and more recently emerging Vision Transformers [3, 37] present highly promising approaches for Precision-Agriculture tasks such as segmentation, fine-grained classi-

fication as well as plant and stem detection. However, their robust application requires an extensive amount of highly variable annotated training data, which is not yet available in this research area. Furthermore, little emphasis is placed on the relevance of combinations of multiple crops and especially weeds, which in our opinion should be exploited to increase detection performance in real-world scenarios.

To address this research need, our work proposes the following contributions:

- We provide a novel large-scale dataset for Precision Agriculture, consisting of highly variable real-world images and multi-modal annotations for a rich set of crop and weed categories.[1]
- We demonstrate its versatility by training and evaluating multiple learning tasks including object detection, stem localization and semantic segmentation.
- We introduce our concept of data ablation to efficiently specialize the dataset to different application requirements. Thereby, we prove our hypothesis that crop detection significantly benefits from incorporating weed classes in addition to the relevant crops.

2. Related work

CNNs have been successfully applied in research domains such as Natural Language Processing [35] and Autonomous Driving [24]. Their structure specialized on pattern recognition enables them to learn rich sets of visual representations of defined image content. Besides advances in network architectures and parallel processing, their increasing performance is facilitated by continuous progress regarding quality and quantity of available datasets. In contrast, for a successful integration into systems for crop and weed manipulation, the scarcity of training and test data as well as the heterogeneity of input modalities still pose a challenge regarding the application of Deep Learning. While there are promising initial approaches demonstrating the potential of CNNs in Precision Agriculture [23, 11, 14], especially using transfer learning [4, 12], the amount and variability of available datasets is still not adequate for their robust and generalized adaptation.

One of the first larger datasets in this area is the Plant Seedlings dataset [10], which focuses on classifying and segmenting 4750 extracted patches of 12 plant species and an additional set of images captured in the wild. However, since most samples were recorded under laboratory conditions, data diversity, especially regarding variations of background and lighting conditions, is limited. A large-scale and well-designed dataset is CropDeep [38], which provides 49k instances for 31 classes of fruits and vegetables annotated for detection and classification and captured in a greenhouse environment for the purpose of automating the picking process. While the data contains a wide range of background variation, the absence of outdoor crop fields and the bias towards ripe fruits limits its applicability for weed intervention systems. Another relevant dataset was published by [31], containing 6 crop and 8 weed species. Besides the laboratory setup, additional image data was recorded at three locations under field conditions resulting in a total of about 8k annotated instances in 1118 images. Furthermore, there are small-scale datasets focusing on specific crop types, such as the DeepWeeds dataset [26], which provides annotations for 8 types of Australian weeds annotated at image level, or the Carrot-Weed dataset [16]. Moreover, a small number of consumer-level plant categorization apps has emerged [22, 33], which however mostly focus on late growth stages. While most contributions rely on RGB image data, there are a few works presenting multi-spectral images [7, 1, 19] as well as synthetic-data generators for crops and weeds [2, 30]. In general, most of these datasets focus on either annotating bounding boxes or semantic masks, but rarely provide both [8], while stem detection is only addressed by few contributions such as [21]. In conclusion, most datasets have strong limitations concerning their application areas, input modalities and annotation types. To the best of our knowledge, there is currently no dataset in the research area of Precision Agriculture comparable to our work regarding the number as well as variability of annotated instances and plant classes for detection, classification, stem localization and semantic labeling.

3. The CropAndWeed dataset

Our proposed dataset provides a vital step towards overcoming the data gap in Precision Agriculture. The image data along with multi-modal annotations is available for academic research and intended to be enhanced in collaboration with the community to gradually increase diversity by adding samples collected all over the world.

3.1. Design considerations

Over a period of four years, we recorded a rich set of image data focusing on a variety of crops and weeds in early growth stages. On the one hand, we collected and annotated samples (*Application Set*) from several hundred commercially used cultivation areas in Austria, which represent realistic validation and test data for practical applications. By incorporating data from diverse locations, this set covers a high variety in soil types as well as typical combinations and distributions of plants. The second part (*Experimental Set*) depicts a wide selection of species specifically grown in a controlled outdoor environment to efficiently enrich the dataset with samples of underrepresented classes.

[1] Images and annotations of the dataset are available for academic use at https://github.com/cropandweed/cropandweed-dataset

Figure 2. Representative selection of all crop and weed varieties included in the dataset.

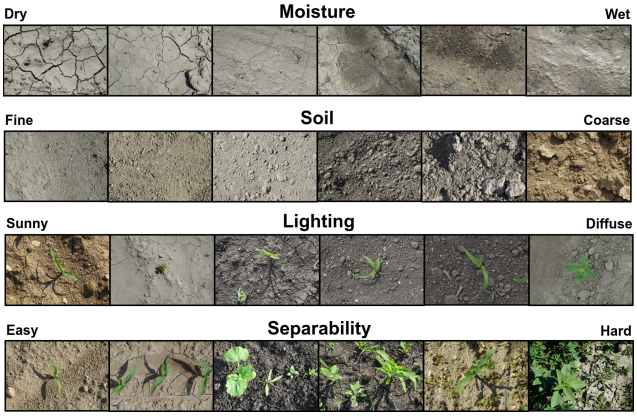

Figure 3. Representative illustrations of parameters defined for each recording session. Categorization is based on the distinctiveness of visual features, such as the presence of cracks for *Moisture* or the hardness of shadows in case of *Lighting*.

Contrary to the former set, these crops and weeds were carefully selected and planted on standardized plots similar to conventional cultivation areas but including only a defined species or combination of plants each. While this approach requires higher manual effort for irrigation and removal of unintended vegetation, it facilitates an unambiguous identification of species even at early growth stages. Furthermore, as these stages are most relevant for tasks such as automatic monitoring and weed removal, cultivating plants for the sole purpose of dataset creation gives us the opportunity to trigger multiple growing periods per season by replacing them significantly earlier than the time they would usually be harvested for commercial purposes.

During data acquisition, special emphasis was placed on achieving a high degree of variability to include most conditions and scenarios realistically encountered during regular farming operations. To cover a broad range of lighting and weather conditions as well as varying soil types, image data was collected over a period of four years between March and July. Each capturing session refers to a unique agricultural site or a specific experimental plot at a specific time and yields an average of only 20 images at intervals of at least three meters to avoid redundancies.

To cover a range of crop and weed species expected to be relevant for current and future applications, we selected a set of 74 plant classes in total, 16 representing species of crops, while the remaining 58 are typically not cultivated intentionally and therefore considered as weeds. A representative composition of instances for these classes illustrating the variability of the dataset regarding species and environmental conditions is presented in Figure 2. The specification furthermore contains the fallback class *Vegetation* for instances which cannot be unambiguously identified in real-world application settings or multi-species experimental plots due to their size ($< 16^2$ pixels) or appearance, since their bounding boxes and segmentation masks are relevant for testing purposes. The full list and description of plant species is available in the supplementary material.

Apart from plant types and their combinations, we aim for a high number of variations in daytime, season, lighting conditions, soil granularity and moisture, as well as different combinations of crop and weed species. Furthermore, the dataset is enriched by negative samples containing no visible vegetation, as well as background clutter in the form of rocks, straw, fertilizer or trash. To the best of our knowledge, we surpass any existing datasets in this domain in terms of variability and flexibility, rendering the CropAndWeed dataset a well-suitable benchmark for future approaches in the area of automated crop and weed manipulation for Precision Agriculture.

3.2. Data acquisition and annotation

Data is collected using a semi-professional SLR with a full-frame sensor to meet our requirements regarding mobility and image quality. All images are manually captured in auto-exposure mode with a constant focal length of 50 mm from an approximate top-down perspective at a height of about 1.1 meters. For each capturing session, we document relevant environmental parameters encountered at the time of recording, as illustrated in Figure 3. They are essential for identifying data gaps during dataset creation and present the opportunity to evaluate the robustness of CNNs against specific environmental conditions. Additionally, the meta-annotation includes time stamps, GPS coordinates as

well as selected camera parameters. All these inputs are exploited to specifically select a representative set of images for fine-grained annotation. Table 1 gives a statistical overview of both subsets.

		Exp	App	Total
Sessions	Recorded	1 363	738	2 101
	Annotated	665	264	929
Images	Recorded	22 597	21 217	43 814
	Annotated	4 990	3 044	8 034
Instances	Annotated	66 877	45 076	111 953

Table 1. Overview of recorded and annotated image data at experimental cultivation plots (*Exp*) and conventional agricultural application sites (*App*).

To increase annotation efficiency, each image is automatically pre-segmented into soil and vegetation pixels. Initially, this was achieved using a traditional method of color-based thresholding [27], which was later replaced by our CNN-based segmentation using preliminary models trained on the annotated data. Thereby, manual effort is reduced to assessing and refining these masks while simultaneously assigning the target classes. Similarly, the resulting mask is used to automatically generate bounding boxes for each plant instance, which are manually refined and enhanced with annotations of each plant's stem position. To ensure consistent annotation quality, data batches with significant ambiguity or high numbers of plants are reviewed by voting of multiple annotators as an additional validation step. The resulting annotations, visualized in Figure 1, provide a thorough foundation for training and evaluating multiple learning tasks to precisely guide plant-intervention systems, such as detection, sub-species classification, semantic and panoptic segmentation as well as anchor-point regression.

4. Methodology

To demonstrate the applicability of our dataset for crop and weed manipulation, we train and benchmark multiple learning tasks. For this purpose, we define multiple dataset variants by mapping the original labels to varying sets of super-classes specialized to different application scenarios.

4.1. Data ablation

The original label specification containing 74 classes is intended to cover a wide range of plants expected to be relevant in current and future applications. While some of them are not yet sufficiently represented for training them separately, they can be combined to super-classes. Since identifying effective mappings requires a certain level of expertise and empirical analysis, this section provides a basis for tailoring the dataset to the cultivated crops and expected weeds at a specific site.

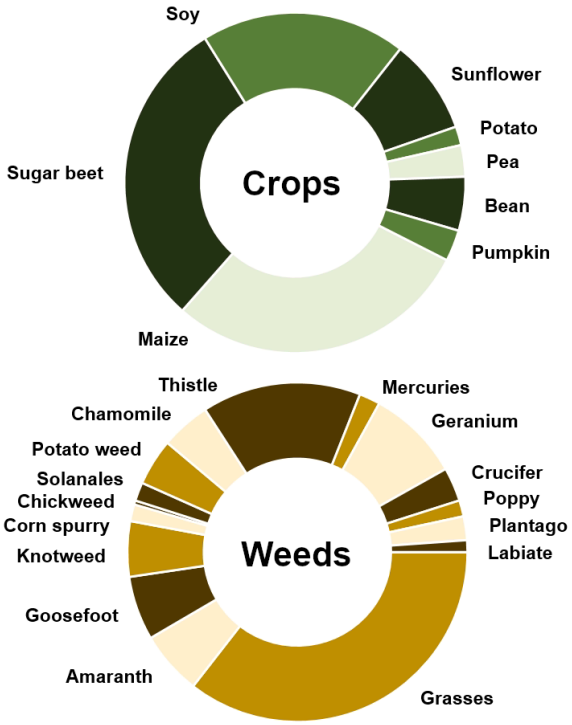

Figure 4. Distribution of crop and weed classes in the *Fine24* dataset variant.

The coarsest possible dataset variant (*Coarse1*) consists of only two classes differentiating between soil and any kind of vegetation, therefore providing pure localization of all available crop and weed species. All other dataset variants presented below provide an additional classification of plant instances at varying levels of granularity and are selected to highlight the resulting flexibility as well as the benefit of incorporating varieties of weeds into the training process. The most fine-grained variant *Fine24* maps the original labels into 8 crop and 16 weed classes based primarily on botanical categorization and, in rare cases, visual similarity. The distribution of annotated object instances and their sizes for each class of this base variant is shown in Figures 4 and 5. In total, crops contribute about 18.7% of all instances and weeds 51.7%, while the remainder belongs to the *Vegetation* class, which holds all *Tiny* instances of the dataset. The distribution of crop and weed classes is closely related to their occurrence frequency in real-world conditions, since a large part of the data was captured at commercially used cultivation areas. However, as discussed in Section 3.1, strongly underrepresented weed classes are additionally grown for the creation of this dataset, thereby ensuring sufficient representation for training and partially mitigating the long-tailed characteristic of the dataset. In general, weeds tend to be more frequent but smaller than crops, rendering them the most challenging target for annotation as well as training, especially regarding class assignment.

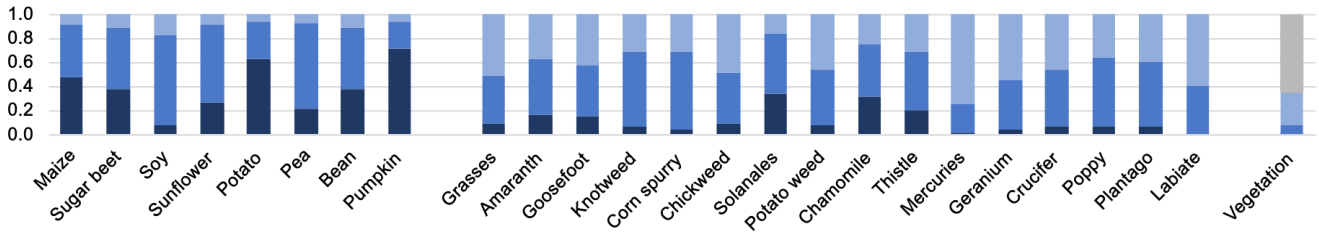

Figure 5. Normalized size distribution of crop and weed instances in the *Fine24* dataset variant clustered by bounding box area in pixels for an image size of 1920x1088 pixels: **Large** $> 128^2 >$ **Medium** $> 32^2 >$ **Small** $> 16^2 >$ **Tiny**.

The presented classes can be condensed into other variants by mapping all types of weeds into one super-class (*CropsOrWeed9*) and furthermore using only a single class for crops as well (*CropOrWeed2*). Additionally, we included variants containing each crop either as a single class or in combination with all other samples mapped into one super-class and denoted these ensembles of multiple models as *Crop1* and *Crop2*, respectively. The exact mappings of all annotated labels for each of the resulting 20 dataset variants are provided in the supplementary material.

4.2. Learning tasks

The defined dataset variants are used to train and compare specialized models for multiple learning tasks and application scenarios. For all experiments, the same constant randomly-generated split between training, validation and test data of 70:15:15 is applied to ensure independence even when combining learning tasks. Training is conducted on a system with two NVIDIA RTX 3090 GPUs. Throughout all experiments, we use stochastic gradient descent [15] for optimization, since we experienced unstable convergence with Adam when using PyTorch [28]. Standard data augmentation techniques such as random scaling and cropping or normalization are applied and the model with best validation performance is selected for evaluation.

Detection For object detection, we experimented with both RetinaNet [17] and YOLOv5 [13] (release 6.0) architectures on the *Fine24* dataset variant with the most relevant results presented in Table 2. Due to its superior performance, we chose YOLOv5l6 for conducting the benchmark experiments for all dataset variants and selected the input size of 1280x1280 pixels, since increasing the dimensions does not yield a significant performance gain considering the increase in training time. After tuning the hyperparameters, we settled on a linear learning rate of 0.1 and a batch size of 16. We use the pre-trained models provided by [13] for initialization and train each variant for 50 epochs.

Segmentation Segmentation experiments are performed for three coarse-grained variants of the dataset. Since segmentation is intended to be combined with detection in real-time for practical applications, we selected an efficient DLA-34 model inspired by [36]. We train from scratch with a batch size of 4, a crop size of 672 pixels, a learning rate of 0.001 and choose the best model after 30 epochs.

	YOLOv5s	YOLOv5m	YOLOv5l
1280²	47.4	51.7	54.6
1920²	51.6	52.3	**56.4**

Table 2. Detection results as AP by model architecture and input resolution in pixels on the validation set of the *Fine24* variant.

Stem Localization We formulated stem localization as a detection problem, for which we applied a pre-trained SSD-300 single shot detector [20] with a VGG-16 [29] backbone for initialization. We used the *Coarse1* variant of the dataset for training, since the results are intended to be combined with our detection module in real-time and therefore do not require redundant fine-grained classification. To capture the context around each stem location, we sample patches of constant size around the annotated anchor points. After experimenting with multiple configurations, we settled on a patch size of 81x81 pixels, as well as an input image size of 300x300 pixels. We trained with a batch size of 80 images and an initial learning rate of 0.0001, which is reduced by a factor of 10 before the 15$^{\text{th}}$ and 30$^{\text{th}}$ epochs. The best model was selected based on validation results after 50 epochs.

5. Evaluation

To establish a benchmark for the applicability of our dataset in different scenarios, we conduct experiments on all dataset variants defined in Section 4.1 for multiple learning tasks and evaluate their results based on established metrics. The test set representing 15% of all images is completely independent of the training and validation data for all experiments.

5.1. Detection

Detection models are trained on each dataset variant and evaluated on the entire test set. As a performance metric, we report the established average precision (*AP*) as de-

scribed by [18], which relies on class-specific intersection-over-union (*IoU*) measures. As opposed to using a single threshold for identifying correct matches, as applied for instance by Pascal VOC [6], this more challenging metric calculates an average of the results for 10 *IoU* thresholds in the range of 0.5 to 0.95. While instances of the *Vegetation* class are not explicitly included for training the models, they are part of the test set and used to ignore any detections matching them during evaluation. This is based on the idea that detected plants which cannot be classified by human annotators should count as neither false nor true positives, as their correctness cannot be ensured. For the same reason, we only evaluate detections larger than 16^2 pixels, which is the minimum size for assigning classes during annotation. Table 3 shows the overall performance of the models trained on each dataset variant, as well as their scores after filtering the test data and detection results by instance size.

	S	M	L	Overall
Fine24	33.1	57.4	72.9	**55.4**
CropsOrWeed9	32.2	71.6	87.0	**71.5**
Crop2	36.1	67.4	84.8	**63.8**
Crop1	18.5	47.9	65.1	**57.3**
CropOrWeed2	33.0	61.1	82.4	**60.7**
Coarse1	29.7	59.0	84.2	**55.2**
	30.4	**60.7**	**79.4**	**60.6**

Table 3. Detection performance (AP) by instance size of models trained on each dataset variant, with *Crop[2/1]* denoting the average of the respective variants for each of the 8 crop classes. Size thresholds are analogous to Figure 5.

Not surprisingly, detection performance significantly increases with the size of instances. However, small instances with sizes between 16^2 and 32^2 pixels can already be classified with promising robustness, while tiny instances below this threshold were previously assigned to the *Vegetation* class as described in Section 3.1 and therefore not included in the training data. Overall detection performance is lowest for the most coarse-grained *Coarse1* and the most fine-grained *Fine24* variants, followed by the 8 individual models trained on only one crop species each with their average denoted as *Crop1*. Between these edge cases, performance increases with the number of incorporated classes. However, since the identification of cultivated crop types is usually more relevant for practical applications than the differentiation between individual types of weeds included in these overall measures, Table 4 provides more detailed per-class detection results.

The model trained and evaluated on the *Coarse1* variant of the dataset addresses the most basic task, since its aim is solely to differentiate any kind of plant annotated in the dataset from the background. Therefore, it provides only a single overall AP value and limited applicability in practice. A slightly more difficult task is presented by *CropOrWeed2* which differentiates between crops and weeds as two opaque classes. Despite their visual similarity, they are still separated by the model with high accuracy. However, practical applications in Precision Agriculture usually require a more fine-grained identification of specific crop types. This feature is addressed by the rows labeled *Crop1* and *Crop2*, which represent the performance of multiple models trained to identify one type of crop each. In the case of the former variants, training data exclusively contains instances of the relevant class, representing the most straight-forward approach to the learning task. Therefore, the models perform rather poorly when they are confronted with samples of other plants in the test data by producing a high number of false-positive detections. The *Crop2* models, on the other hand, mitigate this issue by incorporating a second class consisting of the samples of all remaining plant instances and thereby in most cases significantly improving the scores for the respective crop.

However, evaluating the *CropsOrWeed9* model clearly shows that the best results can be achieved by training all relevant crop classes jointly instead of specializing separate models to each one, which results in a more fine-grained differentiation between the individual crops and the single class holding all weed types. Additionally dissolving the weed class into separate species, as in the *Fine24* dataset variant, can improve scores even further for more visually distinct crops as visible for the *Bean* and *Pumpkin* class. Furthermore, this variant increases the discriminating capacity for weed types, considering that their score now represents an average AP over 15 weed species as opposed to the single class in *CropsOrWeed9*.

In conclusion, the results clearly support our initial hypothesis that the incorporation of other species, especially weeds, increases the performance of models for identifying specific types of crops.

5.2. Segmentation

Segmentation experiments are evaluated using mean intersection over union (mIoU) as introduced by [5]. Table 5 shows the results for all selected dataset variants.

Segmentation performance drastically increases with inverse proportion to the number of classes. The overall differentiation of plant instances from soil areas works similarly well for all models, as visible in the score for the *Soil* class. Assigning individual categories of crops and weeds, however, seems to represent a more challenging task due to their visual similarity and under-representation of individual species compared to soil pixels. This effect is especially visible for the *CropsOrWeed9* variant, where individual crop species produce extremely low IoUs, while sufficiently represented classes perform reasonably well with IoUs of up to 74.1% (*Maize*), resulting in the averaged

	Maize	Sugar beet	Bean	Pea	Sunflower	Soy	Potato	Pumpkin	Weed
Fine24	76.5	79.0	**82.7**	66.9	80.6	53.9	70.8	**89.2**	45.6
CropsOrWeed9	76.1	**81.1**	73.6	**67.5**	**81.4**	**55.0**	**74.6**	88.7	45.7
Crop2	**77.6**	79.5	75.1	67.3	78.5	54.1	65.4	87.5	-
Crop1	75.0	74.4	68.3	29.4	62.9	52.8	21.2	74.3	-
CropOrWeed2					72.6				48.8

Table 4. Detection performance (AP) of models trained on each dataset variant, including results for each crop class, as well as the average of weed classes, where applicable. Note that the lines for *Crop[2/1]* consolidate the scores of all respective crop-specific variants, which only yield results for one specific class each.

	Crop	Weed	Soil	Overall
CropsOrWeed9	19.9	47.6	99.5	55.7
CropOrWeed2	70.0	27.1	99.5	65.5
Coarse1		76.0	99.4	87.7

Table 5. Segmentation performance on test sets as per-class IoU and overall mIoU with the Crop score averaged across all 8 crop species for *CropOrWeed9*.

value of 19.9% in total. The best results are achieved by the *Coarse1* variant, which solely differentiates any kind of plant from the background soil. Apart from providing a promising first benchmark for semantic segmentation, the resulting models can directly be used for pre-segmenting plant instances prior to the annotation stage, as described in Section 3.2. Furthermore, they are combined with the results of detection models to delineate the exact shape of detected instances.

5.3. Stem Localization

For practical applications, the stem-localization module is intended to be combined with object-detection results by generating a fixed number of hypothesis for each frame which are then pruned to select one for each detected bounding box based on their scores. Therefore, the evaluation was conducted on the test set of *Coarse1* with the ground-truth bounding boxes simulating perfect detection results to evaluate pure stem-localization performance. Since the entire dataset contains an average of 13.9 instances per image, the number of stem hypothesis was set to 100 to ensure sufficient results even for images containing high numbers of plants.

Using this setup, the model correctly assigns stem points to an average of 73.1% of samples across all test images. The mean euclidean distances between the assigned locations and corresponding annotations across all frames is 29.1 pixels or 25.5% normalized to the respective bounding-box sizes. Considering the variety of stem appearances, the detection rate seems quite promising using this extremely efficient model. Nevertheless, the approach would benefit from incorporating multiple scales to improve localization accuracy across the inherently broad range of plant sizes.

5.4. Discussion

Training and evaluating detection models on all dataset variants specified in Section 4.1 clearly showed that best results for identifying individual crop types are achieved by training them jointly and in combination with a single or multiple classes of weeds, as opposed to creating models specialized to a single crop. The worst performance results from omitting weeds altogether in the training process. Therefore, our dataset is well suitable for specializing models to the requirements of specific applications in the research area of Precision Agriculture by creating specialized dataset variants including all relevant crops and weeds expected to be encountered at the respective site.

Furthermore, the multi-modal annotations facilitate efficient combinations of multiple learning tasks as well as dataset variants. Detection can be enriched by incorporating the results of segmentation and stem localization to provide even more thorough scene understanding. We demonstrate this approach by applying our binary-segmentation and stem-localization models to the output of the *Fine24* detection experiments, as presented in Figure 6.

Combining detection with these modalities provides a more accurate pixel-wise segmentation of each detected plant instance with the corresponding label extracted from the detection result. Moreover, stem localization is performed inside the detected bounding boxes, resulting in a rich set of information about each frame which is applicable for multiple tasks in the field of Precision Agriculture.

6. Conclusion

We introduced a novel large-scale multi-modal dataset for crop and weed manipulation in the context of Precision Agriculture consisting of more than 8k high-quality images and about 112k annotated plant instances. In addition to bounding boxes, segmentation masks and stem positions, annotations include a fine-grained classification into 16 crop and 58 weed species, as well as extensive meta-annotations of relevant environmental and recording parameters. Special emphasis was placed on high data variability and the representation of rare types of weeds, which were specifically cultivated for the purpose of dataset creation.

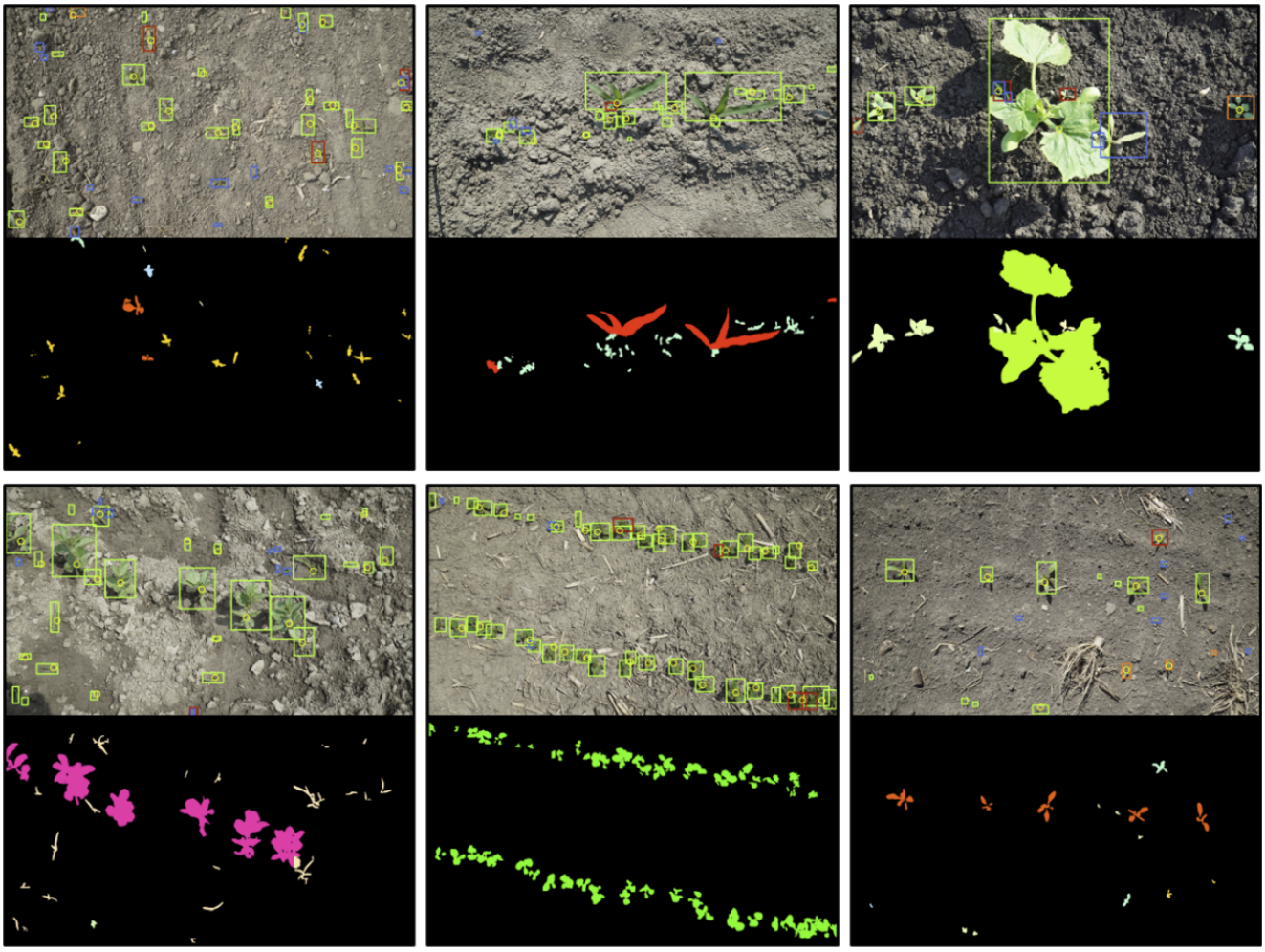

Figure 6. Representative selection of combined qualitative results on the test set for detection, segmentation and stem localization. Top rows show detection results of the *Fine24* model (green: correct detection (*TP*), orange: correct localization, but incorrect class, red: incorrect detection (*FP*), blue: undetected ground-truth object (*FN*). The detections are used as inputs for stem localization (yellow) and the *Coarse1* segmentation model. Results of the latter are demonstrated in the bottom rows, with colors corresponding to the detected classes.

We demonstrated the resulting flexibility and versatility by training and benchmarking CNNs for the tasks of object detection, stem localization and segmentation on multiple variants of the dataset and performing a thorough evaluation regarding different crop species as well as instance sizes. Based on the results we were able to showcase how our approach can be tailored to specific application scenarios and prove our hypothesis that incorporating multiple weed species into the training dataset improves detection performance for the relevant crops. Furthermore, we showcased the potential of flexibly combining multiple learning tasks trained on variants of the dataset.

While the results presented in this work are highly promising, we plan to further enrich the dataset in cooperation with the research community to increase data variability with samples from locations all over the world and additional plant classes. Furthermore, we intend to add multi-object tracking to increase robustness by re-identifying individual plant instances over time. This implies the incorporation of different input modalities such as continuous image sequences captured by UAVs and other mobile platforms to increase recording efficiency and further facilitate the applicability for a wide range of applications.

Acknowledgement. We would like to thank the Federal Ministry for Climate Action, Environment, Energy, Mobility, Innovation and Technology, and the Austrian Research Promotion Agency (FFG) for co-financing the "ICT of the Future" research project HARV.EST (FFG No. 892347). Additionally, we would like to thank Caroline Huber, Reinhard Neugschwandtner and Helmut Wagentristl at the experimental farm Groß-Enzersdorf and our annotation team consisting of Gulnar Bakytzhan, Vanessa Klugsberger and Marlene Glawischnig.

References

[1] Nived Chebrolu, Philipp Lottes, Alexander Schaefer, Wera Winterhalter, Wolfram Burgard, and Cyrill Stachniss. Agricultural robot dataset for plant classification, localization and mapping on sugar beet fields. *The International Journal of Robotics Research*, 36(10):1045–1052, 2017.

[2] Maurilio Di Cicco, Ciro Potena, Giorgio Grisetti, and Alberto Pretto. Automatic model based dataset generation for fast and accurate crop and weeds detection. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5188–5195. IEEE, 2017.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[4] Borja Espejo-Garcia, Nikos Mylonas, Loukas Athanasakos, Spyros Fountas, and Ioannis Vasilakoglou. Towards weeds identification assistance through transfer learning. *Computers and Electronics in Agriculture*, 171:105306, 2020.

[5] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015.

[6] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[7] Adnan Farooq, Jiankun Hu, and Xiuping Jia. Analysis of spectral bands and spatial resolutions for weed classification via deep convolutional neural network. *IEEE Geoscience and Remote Sensing Letters*, 16(2):183–187, 2018.

[8] Mulham Fawakherji, Ali Youssef, Domenico Bloisi, Alberto Pretto, and Daniele Nardi. Crop and weeds classification for precision agriculture using context-independent pixel-wise segmentation. In *2019 Third IEEE International Conference on Robotic Computing (IRC)*, pages 146–152. IEEE, 2019.

[9] Leonard P Gianessi. The increasing importance of herbicides in worldwide crop production. *Pest Management Science*, 69(10):1099–1105, 2013.

[10] Thomas Mosgaard Giselsson, Rasmus Nyholm Jørgensen, Peter Kryger Jensen, Mads Dyrmann, and Henrik Skov Midtiby. A public image database for benchmark of plant seedling classification algorithms. *arXiv preprint arXiv:1711.05458*, 2017.

[11] Honghua Jiang, Chuanyin Zhang, Yongliang Qiao, Zhao Zhang, Wenjing Zhang, and Changqing Song. Cnn feature based graph convolutional network for weed and crop recognition in smart farming. *Computers and Electronics in Agriculture*, 174:105450, 2020.

[12] Zichao Jiang. A novel crop weed recognition method based on transfer learning from vgg16 implemented by keras. In *IOP Conference Series: Materials Science and Engineering*, volume 677/3, page 032073. IOP Publishing, 2019.

[13] Glenn Jocher, K Nishimura, T Mineeva, and R Vilariño. yolov5. *Code repository https://github.com/ultralytics/yolov5*, 2020.

[14] Abbas Khan, Talha Ilyas, Muhammad Umraiz, Zubaer Ibna Mannan, and Hyongsuk Kim. Ced-net: crops and weeds segmentation for smart farming using a small cascaded encoder-decoder architecture. *Electronics*, 9(10):1602, 2020.

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[16] Petre Lameski, Eftim Zdravevski, Vladimir Trajkovik, and Andrea Kulakov. Weed detection dataset with rgb images taken under variable light conditions. In *International Conference on ICT Innovations*, pages 112–119. Springer, 2017.

[17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017.

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[19] Qinghui Liu, Michael C Kampffmeyer, Robert Jenssen, and Arnt-Borre Salberg. Multi-view self-constructing graph convolutional networks with adaptive class weighting loss for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 44–45, 2020.

[20] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*, pages 21–37. Springer, 2016.

[21] Philipp Lottes, Jens Behley, Nived Chebrolu, Andres Milioto, and Cyrill Stachniss. Joint stem detection and crop-weed classification for plant-specific treatment in precision farming. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8233–8238. IEEE, 2018.

[22] Patrick Mäder, David Boho, Michael Rzanny, Marco Seeland, Hans Christian Wittich, Alice Deggelmann, and Jana Wäldchen. The flora incognita app–interactive plant species identification. *Methods in Ecology and Evolution*, 12(7):1335–1342, 2021.

[23] Andres Milioto, Philipp Lottes, and Cyrill Stachniss. Real-time semantic segmentation of crop and weed for precision agriculture robots leveraging background knowledge in cnns. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2229–2235. IEEE, 2018.

[24] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4990–4999, 2017.

[25] E-C Oerke. Crop losses to pests. *The Journal of Agricultural Science*, 144(1):31–43, 2006.

[26] Alex Olsen, Dmitry A Konovalov, Bronson Philippa, Peter Ridd, Jake C Wood, Jamie Johns, Wesley Banks, Benjamin

Girgenti, Owen Kenny, James Whinney, et al. Deepweeds: A multiclass weed species image dataset for deep learning. *Scientific Reports*, 9(1):1–12, 2019.

[27] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979.

[28] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8026–8037, 2019.

[29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[30] Soren Skovsen, Mads Dyrmann, Anders K Mortensen, Morten S Laursen, René Gislum, Jorgen Eriksen, Sadaf Farkhani, Henrik Karstoft, and Rasmus N Jorgensen. The grassclover image dataset for semantic and hierarchical species understanding in agriculture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[31] Kaspars Sudars, Janis Jasko, Ivars Namatevs, Liva Ozola, and Niks Badaukis. Dataset of annotated food crops and weed images for robotic computer vision control. *Data in Brief*, page 105833, 2020.

[32] K Thenmozhi and U Srinivasulu Reddy. Crop pest classification based on deep convolutional neural network and transfer learning. *Computers and Electronics in Agriculture*, 164:104906, 2019.

[33] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8769–8778, 2018.

[34] Yonghua Xiong, Longfei Liang, Lin Wang, Jinhua She, and Min Wu. Identification of cash crop diseases using automatic image segmentation algorithm and deep learning with expanded dataset. *Computers and Electronics in Agriculture*, 177:105712, 2020.

[35] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine*, 13(3):55–75, 2018.

[36] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2403–2412, 2018.

[37] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6881–6890, 2021.

[38] Yang-Yang Zheng, Jian-Lei Kong, Xue-Bo Jin, Xiao-Yi Wang, Ting-Li Su, and Min Zuo. Cropdeep: The crop vision dataset for deep-learning-based classification and detection in precision agriculture. *Sensors*, 19(5):1058, 2019.