WeedsGalore: A Multispectral and Multitemporal UAV-based Dataset for Crop and Weed Segmentation in Agricultural Maize Fields Supplementary Material

Ekin Celikkan[1,2] Timo Kunzmann[1] Yertay Yeskaliyev[1] Sibylle Itzerott[1]
Nadja Klein[3] Martin Herold[1]
[1]GFZ German Research Centre for Geosciences [2]Humboldt-Universität zu Berlin
[3]Scientific Computing Center, Karlsruhe Institute of Technology
{ekin.celikkan, itzerott, herold}@gfz-potsdam.de nadja.klein@kit.edu

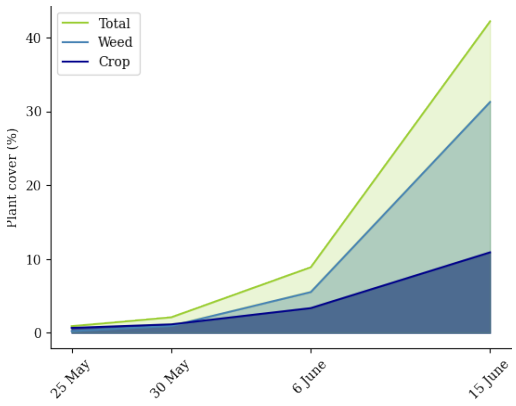

Figure 1. Pixel-level plant cover over time.

This supplementary material provides additional details and insights for the WeedsGalore dataset. Sec. 1 describes the steps for orthomosaic generation, Sec. 2 shows the percentage plant cover change through the data collection period, Sec. 3 includes examples from the dataset, and Sec. 4 shows qualitative examples supporting the analyses done in the main paper. Sec. 5 describes the theoretical grounds of the dropout approximation of Bayesian Variational Inference, and Sec. 6 provides further results on cross dataset evaluation, as well as more details on Maize2024 data. Lastly, Sec. 7 provides runtime estimates and number of parameters.

1. Orthomosaic Generation

The single-band images are processed into an orthomosaic with the software Agisoft Metashape using default parameter values recommended for DJI Phantom 4 Multispectral data processing [1]. Prior to alignment, in order to get higher coordinate accuracy for the raw images, they are processed with Post-Processing Kinematics (PPK) for positional correction. We used reference data provided by the Satellite Positioning Service of the German National Survey (SAPOS) [2].

2. Multitemporal Overview

Total per-pixel plant cover change through the acquisition period can be seen from Fig. 1 for both crops and weeds. It shows a steep increase over time. This results in a large variety in terms of plant size in our dataset. It is also seen that most of the plant cover comes from weeds, instead of crops. This again points out to the diversity in our dataset.

3. Visual Examples

More visual examples from the dataset, including single band images and semantic masks can can be seen from Fig. 2.

4. Qualitative Examples and Failure Cases

4.1. Semantic Segmentation

Quickweed (*Galinsoga parviflora*) is often misclassified as amaranth (*Amaranthus retroflexus*). There are two possible reasons behind this. Firstly, amaranth is the most represented weed class in the dataset while quickweed is the least. The second is their appearances. They are both broadleaf plants with similar phenotypes, which makes it challenging to distinguish them from aerial drone imagery, where the resolution is a limiting factor (see Fig. 3).

4.2. Instance Segmentation

Qualitative results for instance segmentation can be seen from Fig. 4. For plants in early growth stages, the objects are very small, hence the task is challenging and some instances are missed. For later dates, the performance is poor

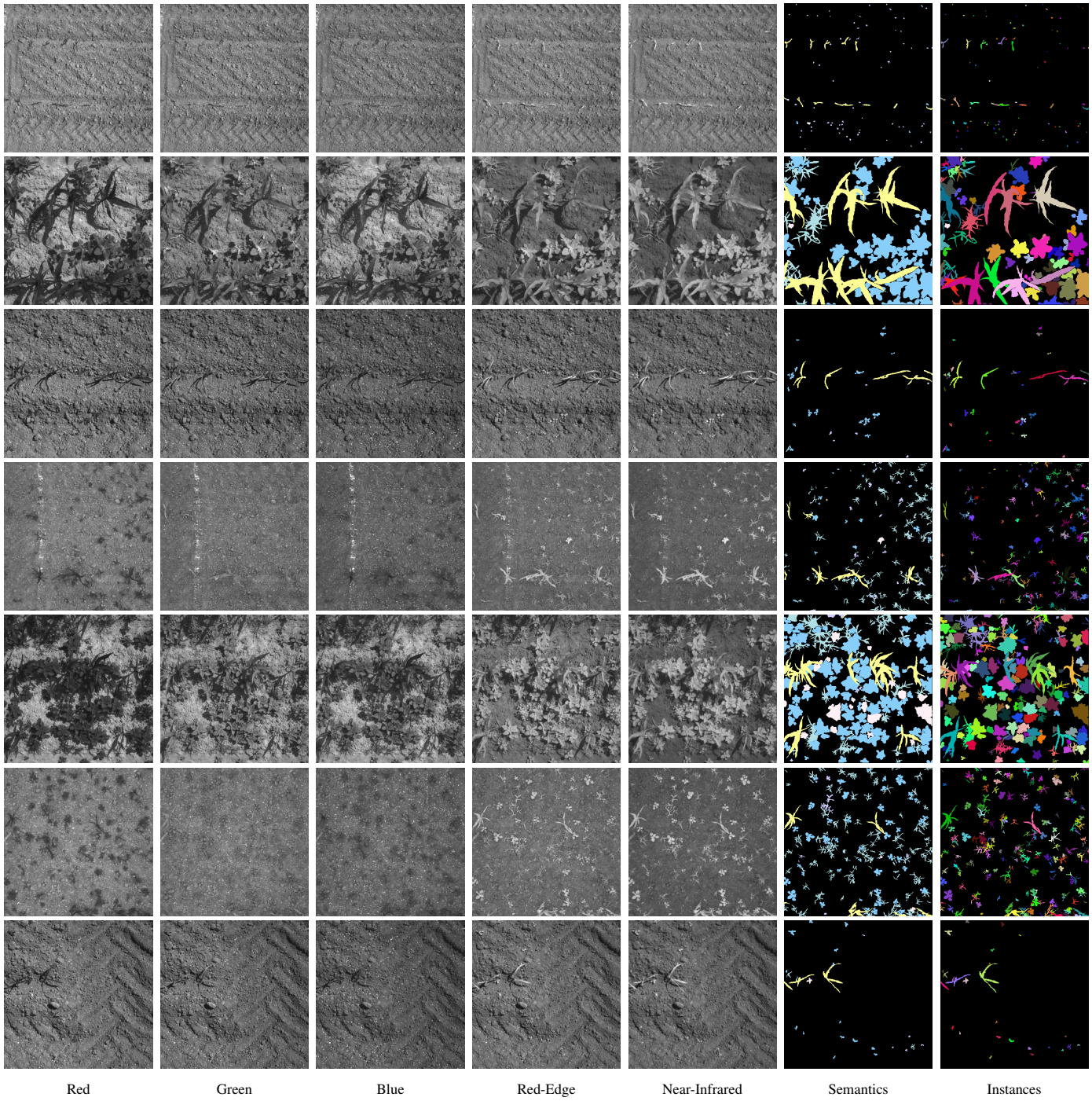

Figure 2. **Examples scenes from WeedsGalore.** Five single-band channels (red, green, blue, red-edge, near-infrared), semantic (maize, amaranth, barnyard grass, quickweed, weed other) and instance masks.

for areas with high overlap (common for large plant cover), and plant organs can be oversegmented into multiple objects (e.g. different leaves).

5. Theoretical Grounds of Variational Inference with MC Dropout

Let D_n be the dataset, on which the model is trained, and x^* be a new input, for which we would like to in-

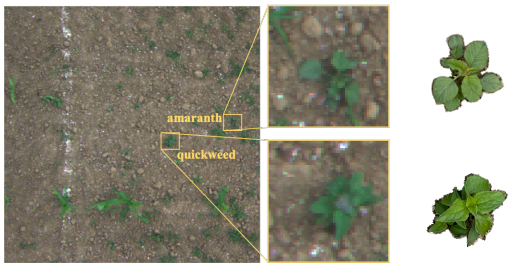

Figure 3. **Amaranth vs. quickweed.** (a) An RGB image from our dataset where (b) two instances of amaranth and quickweed are marked. (c) Examples of the same species from a hand-held camera, where they are more distinguishable. Best viewed on screen zoomed in.

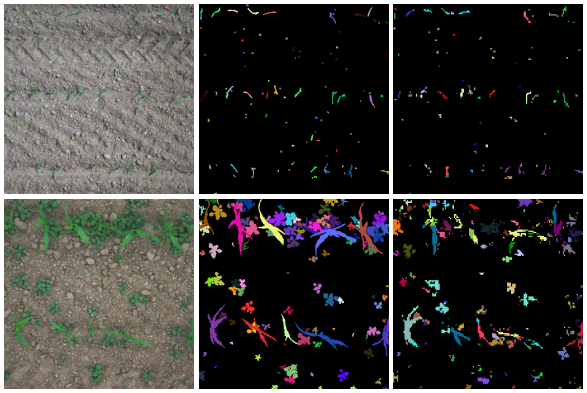

Figure 4. Qualitative results on instance segmentation.

fer the posterior predictive distribution $p(y^*|x^*, D_n) = \int p(y^*|x^*, w)p(w|D_n)\,dw$, where w are the trainable parameters of the network and $p(w|D_n)$ their posterior distribution. Due to its intractability, we approximate the posterior by the variational density q, which minimizes the Kullback-Leibler divergence $\mathrm{KL}(q(w)||p(w|D_n))$. This is equivalent to maximizing the evidence lower bound (ELBO) [5], and can be practically realized by commonly used stochastic optimizers and cost functions (*e.g.* cross-entropy loss) [3]. In VI with MC-Dropout, $q(w) \sim \mathrm{Bernoulli}(p_d)$ is assumed, where p_d is the dropout probability. Different from the traditional approach of using dropout solely for its regularization effect (i.e., to avoid overfitting), the dropout layers are also kept on during test time, hence each prediction is a sample from the approximate posterior. Those samples can be used to calculate the predictive entropy, which is a measure of both epistemic and aleatoric uncertainty [4].

In the case of semantic segmentation, where each pixel is assigned one of the C classes, and K is the number of

Source	UAV	Maize	$\mathrm{IoU}_{\mathrm{bg}}$	$\mathrm{IoU}_{\mathrm{crop}}$	$\mathrm{IoU}_{\mathrm{weed}}$	mIoU
PhenoBench	✓	✗	95.96	21.18	0.61	39.25
CropAndWeed	✗	✓	94.52	0.00	16.92	37.15
MaizeOrWeed	✗	✓	93.73	8.54	9.33	37.20
WeedsGalore train set	✓	✓	97.97	67.93	72.08	79.33

Table 1. Semantic segmentation scores on the WeedsGalore test set for different source domains (i.e. training datasets).

samples drawn from the posterior, the predictive entropy is given as

$$H[y^*|x^*, D_n] = -\sum_{c\in\mathcal{C}} \left(\frac{1}{K}\sum_{k=1}^{K} p(y^* = c|\mathrm{x}^*, w_k) \cdot \log\left(\frac{1}{K}\sum_{k=1}^{K} p(y^* = c|\mathrm{x}^*, w_k) \right) \right).$$

6. Cross-Dataset Evaluation

6.1. Results on WeedsGalore

In this section, we provide scores on WeedsGalore test set where the network is trained on different datasets, namely PhenoBench [7], CropAndWeed [6], and CropAndWeed (1753 image subset of CropAndWeed, containing only maize scenes). The quantitative results are shown in Tab. 1. Phenobench [7] achieves 21.18% for crops, but completely misses weeds. CropAndWeed [6] variants perform better for weeds, but again poorly on crops, showing that even a dataset for the same crop does not generalize to another acquisition mode. To sum up, the models trained on other datasets perform poorly on our test set. This outcome is not surprising, yet it strongly indicates that there is a need for specifically tailored data for this application setting.

6.2. Results on Maize2024

6.2.1 Further Dataset Information

The OOD field (Maize2024) is another agricultural site, located in Marquardt, Potsdam, Germany (52°27'51,1" 12°57'35,6") and covers an area of approximately 2550m^2. The crop was sown in May 2024 (i.e. one year later than WeedsGalore) and on June 6, 2024 herbicide was applied to 12 out of 24 patches with a spraying tractor. The patches were chosen taking the field slope, and wind direction (at the time of the herbicide spraying) into account, so that patches with different characteristics are present in both control and test groups. As a result, data was collected with the same drone on June 25, where a considerable amount of weeds was eliminated in the sprayed patches whereas control (i.e. not sprayed) areas had full plant cover (i.e. complete weed infestation).

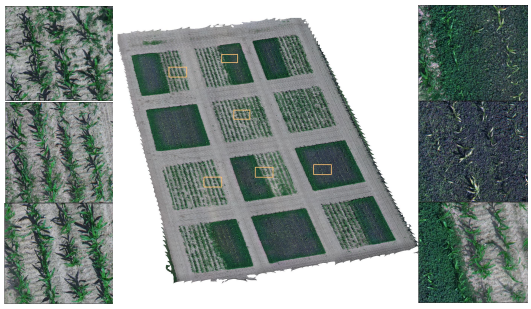

Figure 5. **Annotated reference data for Maize2024**. Each marked area corresponds to approximately 15m^2.

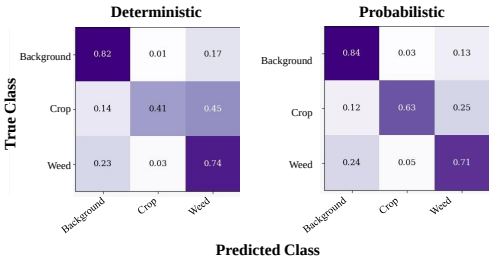

Figure 6. **Normalized confusion matrices for Maize2024.** Results for 3-class semantic segmentation with deterministic (left) vs probabilistic (right) variants of DeepLabv3+, trained on WeedsGalore.

Method	#Parameters	Runtime [ms]
MaskFormer	41M	72.45
DeepLabv3+	39M	103.12
Prob. DeepLabv3+	39M	627.37

Table 2. Comparison of model parameters and inference runtime for different methods on WeedsGalore scenes. Reported scores are computed on a Quadro P4000 GPU, and for 3-channel input, 3 output classes, and 5 forward passes for the probabilistic variant.

6.2.2 Confusion Matrices

The confusion matrices on Maize2024 data for deterministic and probabilistic models can be seen in Fig. 6.

7. Computational Comparison

Tab. 2 provides the number of parameters and runtime of the models that are used in this paper. Compared to the deterministic version, the number of parameters of the probabilistic one using MC dropout is the same. For MC dropout we need to run several forward passes, increasing the runtime accordingly.

References

[1] DJI Phantom 4 Multispectral data processing. https://agisoft.freshdesk.com/support/solutions/articles/31000159853-dji-phantom-4-multispectral-data-processing. [Accessed 17-07-2024]. 1

[2] SAPOS GPPS. https://geobasis-bb.de/lgb/de/geodaten/raumbezug-sapos/sapos-gpps/#. [Accessed 17-07-2024]. 1

[3] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1050–1059. PMLR, 2016. 3

[4] Jishnu Mukhoti and Yarin Gal. Evaluating bayesian deep learning methods for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2018. 3

[5] J. T. Ormerod and M. P. Wand. Explaining variational approximations. *The American Statistician*, 64(2):140–153, 2010. 3

[6] Daniel Steininger, Andreas Trondl, Gerardus Croonen, Julia Simon, and Verena Widhalm. The cropandweed dataset: A multi-modal learning approach for efficient crop and weed manipulation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3729–3738, 2023. 3

[7] Jan Weyler, Federico Magistri, Elias Marks, Yue Linn Chong, Matteo Sodano, Gianmarco Roggiolani, Nived Chebrolu, Cyrill Stachniss, and Jens Behley. PhenoBench: A Large Dataset and Benchmarks for Semantic Image Interpretation in the Agricultural Domain . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 46(12):9583–9594, 2024. 3