ReLeaf: Benchmarking Leaf Segmentation across Domains and Species

Supplementary Material

Robert Martinko[1,2] Daniel Steininger[1] Julia Simon[1] Andreas Trondl[1] Matthias Blaickner[2]
[1]AIT Austrian Institute of Technology, Center for Vision, Automation & Control
[2]University of Applied Sciences Technikum Wien, Computer Science & Applied Mathematics
{robert.martinko,daniel.steininger,julia.simon,andreas.trondl.fl}@ait.ac.at
matthias.blaickner@technikum-wien.at

This supplementary document complements the main paper with additional statistics regarding our benchmark dataset and more detailed quantitative and qualitative experimental results to facilitate a more thorough understanding of applied methods and insights.

A. Benchmark dataset

Our *CropAndWeedAndLeaf* benchmark comprises a representative set of images sampled from 23 classes of the *CropAndWeed* dataset [5]. Tab. 1 gives a complete list of the exact species and the *ReLeaf* model's performance on each of them. It shows strong variations between classes related to their appearance and leaf structure, highlighting the overall need for more variable leaf-segmentation data.

B. Extended results

The following sections show detailed results of both our ablation and cross-dataset experiments.

B.1. Architecture ablation

As stated in the main paper, RF-DETR was omitted from the final ablation after initial experiments showed drastically inferior performance compared to other architectures under the applied hardware constraints. Tab. 2 provides an overview of these initial results. While inference times of RF-DETR are comparable to those of YOLO26 and even slightly lower than for Detectron2, mAP values are significantly inferior to both. At the same time, training times are higher by a factor of 3.5 compared to Detectron2 at the same image resolution and 6.3 compared to YOLO26 models trained at even higher resolutions, which are not available for RF-DETR due to its substantial VRAM requirements. Mitigating these discrepancies and making transformer-based architectures usable for our hardware and data setup could still be achieved by using specialized architecture variants (e.g., RT-DETR [3], Deformable DETR [9]) or dedicated training strategies, such as memory-efficient attention mechanisms or staged training starting from plant-level instance-segmentation datasets. These options may be worth exploring in future extensions of this work, but were

Table 1. Overview of 23 plant species included in *CropAndWeedAndLeaf* with leaf-segmentation annotations for 15 samples each, along with corresponding instance-segmentation performance (% $mAP^{50\text{-}95}$). Scores are results of the selected model architecture (YOLO26 *Medium*, input size of 768^2 pixels), trained on the combined *ReLeaf* dataset.

Species name	***Botanical name***	**Score**
Black-bindweed	*Fallopia convolvulus*	41.6
Broad bean	*Vicia faba*	49.7
Common bean	*Phaseolus vulgaris*	66.5
Common corncockle	*Agrostemma githago*	20.2
Common sunflower	*Helianthus annuus*	73.0
Copse bindweed	*Fallopia dumetorum*	43.9
Cornflower	*Centaurea cyanus*	30.0
Creeping thistle	*Cirisium arvense*	39.9
Field sowthistle	*Sonchus arvensis*	34.7
Maize	*Zea mays*	41.2
Maple-leaf goosefoot	*Chenopodium hybridum*	31.7
Pea	*Pisum sativum*	11.2
Poppy	*Papaver*	54.5
Potato	*Solanum tuberosum*	32.5
Red-root amaranth	*Amaranthus retroflexus*	57.9
Redshank	*Persicaria maculosa*	26.7
Ribwort plantain	*Plantago lanceolata*	34.8
Soybean	*Glycine max*	27.2
Squash	*Cucurbita*	60.9
Sugar beet	*Beta vulgaris s. vulgaris*	66.0
Thornapple	*Datura stramonium*	57.6
White goosefoot	*Chenopodium album*	39.4
Zucchini	*Cucurbita pepo var. gir.*	68.2

Table 2. Initial ablation results for RF-DETR *SegPreview* with multiple input resolutions, including mask accuracy as well as training and inference times (training at an input resolution of 768^2 pixels was not feasible due to high VRAM requirements). Results of selected YOLO26 and Detectron2 variants (*Medium* and *ResNet-101*, respectively) are provided for reference. All models are trained for 20 epochs using the *PhenoBench* training/validation splits and evaluated on the corresponding test set.

	Size [px]	Accuracy [% mAP]	Training [h/epoch]	Inference [ms/image]
RF-DETR	192^2	55.9	0.85	18.0
	384^2	59.2	0.94	19.7
	576^2	59.9	1.13	23.4
Detectron2	*576^2*	*80.0*	*0.32*	*33.1*
YOLO26	*768^2*	*78.0*	*0.18*	*21.0*

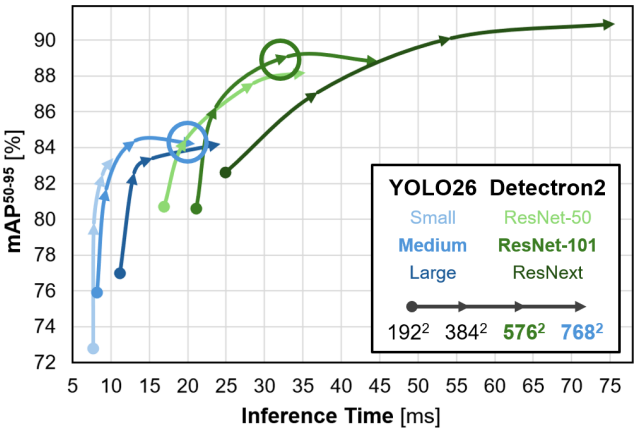

Figure 1. Comparison of bounding-box accuracy (% mAP^{50-95}) and inference time (ms) of leaf-segmentation for multiple model-architecture variants and input resolutions on the *PhenoBench* test set. Selected variants providing a reasonable trade-off for real-time deployment are marked by circles.

not within the scope of our analysis.

In addition to the accuracy of leaf-segmentation masks provided in Fig. 4 of the main paper, we present the corresponding bounding-box accuracy in Fig. 1 to allow a differentiation between the models' capabilities of correctly localizing individual leaves and accurately delineating their contours. As visible in the results, the discrepancy is significantly larger for Detectron2 [8] compared to YOLO26 [1], indicating that it is a superior choice for applications requiring pure detection in the form of bounding boxes without fine-grained object masks. For the purposes investigated in this work, however, YOLO26 remains a more suitable choice due to its lower processing time. Tab. 3 shows the exact values forming the basis of both charts.

Tab. 4 shows the performance of our selected configurations of YOLO26 and Detectron2 on different leaf sizes. It confirms our prior assumption that two-stage detectors like Detectron2 excel at retrieving and accurately delineating small objects at the cost of higher parameter counts and therefore longer inference times than single-stage variants.

Table 3. Comparison of bounding-box (**Box**) and segmentation-mask (**Mask**) accuracies (% mAP^{50-95}) and inference times (**t**) (ms) of leaf-segmentation for multiple model-architecture variants and input resolutions on the *PhenoBench* test set. Selected variants providing a reasonable trade-off for real-time deployment are highlighted by bold text.

		Size	Box	Mask	t
YOLO26	**Small**	192^2	72.8	65.0	7.7
		384^2	79.8	75.0	7.8
		576^2	82.5	79.1	8.9
		768^2	83.4	80.4	10.3
	Medium	192^2	75.9	66.8	8.2
		384^2	81.7	76.6	9.4
		576^2	84.4	80.4	13.2
		768^2	**84.2**	**81.4**	**21.0**
	Large	192^2	77.0	67.5	11.2
		384^2	82.6	76.9	13.1
		576^2	83.4	79.7	15.5
		768^2	84.2	81.3	24.3
Detectron2	**ResNet-50**	192^2	80.7	75.9	16.9
		384^2	84.5	79.2	19.9
		576^2	87.6	81.5	28.6
		768^2	88.2	81.5	35.3
	ResNet-101	192^2	80.6	75.9	21.1
		384^2	86.3	80.4	23.7
		576^2	**89.1**	**82.4**	**33.1**
		768^2	88.8	82.1	44.9
	ResNext	192^2	82.6	77.4	24.9
		384^2	87.1	80.9	37.0
		576^2	90.1	83.7	54.3
		768^2	90.9	83.9	75.7

Table 4. Comparison of bounding-box (**Box**) and segmentation-mask (**Mask**) accuracies (mAP^{50-95}) for varying objects sizes (**S**mall $< 32^2$ pixels $<$ **M**edium $< 96^2$ pixels $<$ **L**arge) using the selected configurations of YOLO26 (*Medium* with an input resolution of 768^2 pixels) and Detectron2 (*ResNet-101* with an input resolution of 576^2 pixels).

	Box			Mask		
	S	**M**	**L**	**S**	**M**	**L**
YOLO26	69.9	82.2	91.3	69.5	79.3	89.1
Detectron2	87.4	88.0	91.7	80.9	80.9	86.3

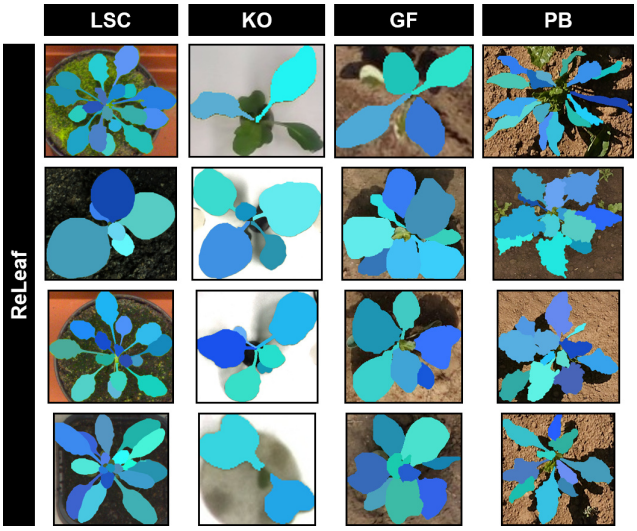

Figure 2. Representative leaf-segmentation results of selected YOLO26 configuration (*Medium* with an input resolution of 768^2 pixels) trained on a combination of *LSC* [4], *Komatsuna* (KO) [6], *GrowliFlower* (GF) [2] and *PhenoBench* (PB) [7] datasets and evaluated on the respective test sets.

B.2. Cross-dataset experiments

Fig. 2 visualizes our multi-domain *ReLeaf* model's results on a selection of representative test images from *LSC* [4], *Komatsuna* [6], *GrowliFlower* [2] and *PhenoBench* [7], confirming its stable performance across both species and environmental conditions.

Figs. 3 and 4 show the results of models trained on each dataset and evaluated on one representative image for each of the 23 plant species in the *CropAndWeedAndLeaf* benchmark. Models trained on laboratory data (*LSC*, *Komatsuna*) struggle most to generalize to novel plant species in real-world settings. While the combined *ReLeaf* model occasionally misses small inner leaves, it significantly improves adaptation to the strongly varying visual appearances and leaf configurations of different crop and weed species.

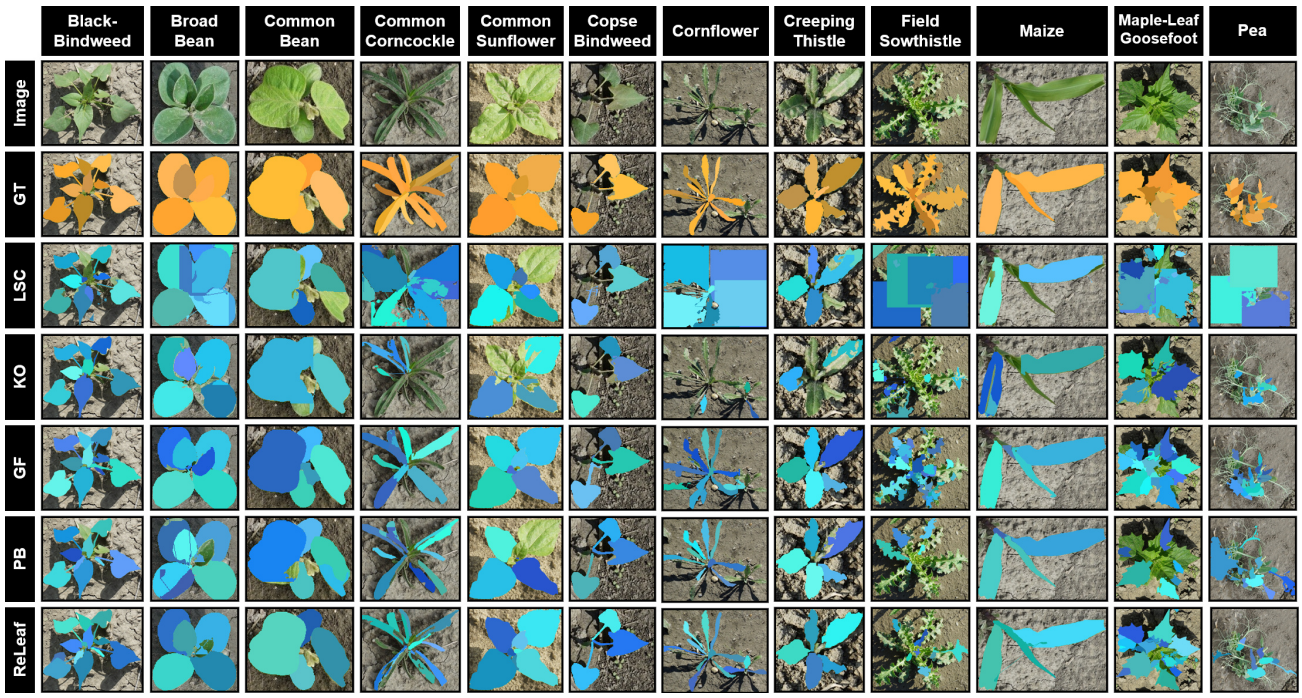

Figure 3. Representative leaf-segmentation results of YOLO26 models (*Medium*, 768^2 pixels input resolution) trained on ***LSC*** [4], ***KO****matsuna* [6], ***G****rowli****F****lower* [2] and ***P****heno****B****ench* [7] datasets and their combination (***ReLeaf***), evaluated on all 23 species of the *CropAndWeedAndLeaf* benchmark, along with corresponding ground truth (**GT**) (Part 1).

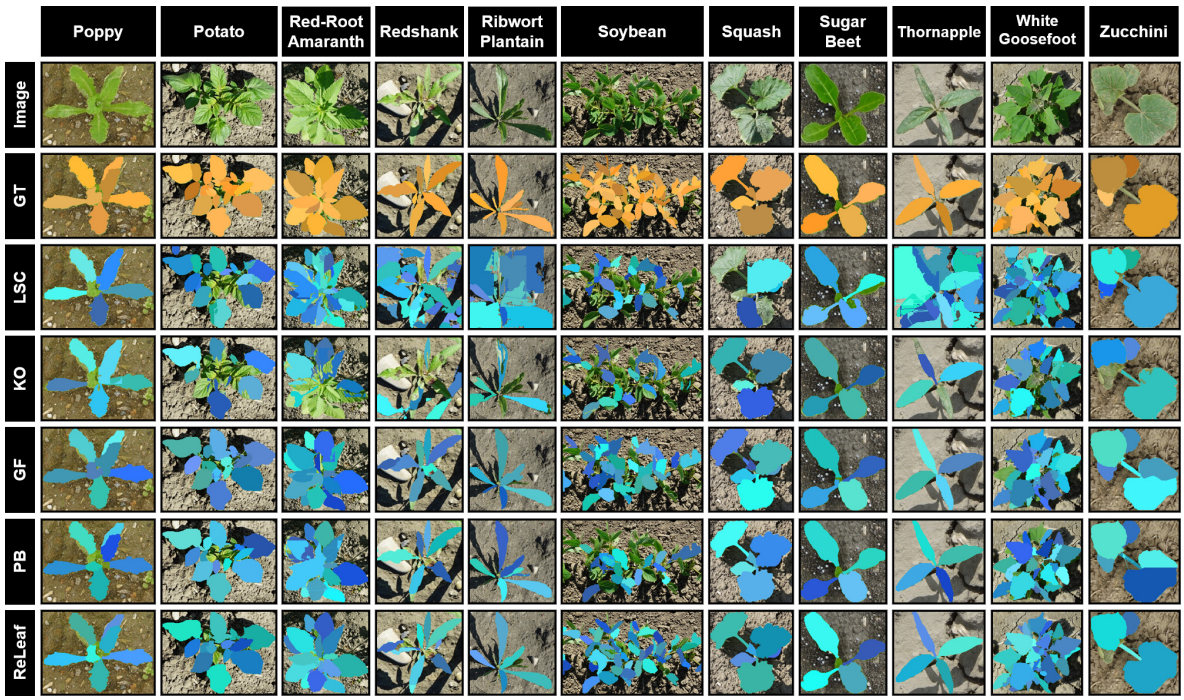

Figure 4. Representative leaf-segmentation results of YOLO26 models (*Medium*, 768^2 pixels input resolution) trained on ***LSC*** [4], ***KO****matsuna* [6], ***G****rowli****F****lower* [2] and ***P****heno****B****ench* [7] datasets and their combination (***ReLeaf***), evaluated on all 23 species of the *CropAndWeedAndLeaf* benchmark, along with corresponding ground truth (**GT**) (Part 2).

References

[1] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolo. https://github.com/ultralytics/ultralytics, 2024. Accessed: 2025-10-11. 2

[2] Jana Kierdorf, Laura Verena Junker-Frohn, Mike Delaney, Mariele Donoso Olave, Andreas Burkart, Hannah Jaenicke, Onno Muller, Uwe Rascher, and Ribana Roscher. Growliflower: An image time-series dataset for growth analysis of cauliflower. *Journal of Field Robotics*, 40(2):173–192, 2023. 3, 4

[3] Wenyu Lv, Yian Zhao, Qinyao Chang, Kui Huang, Guanzhong Wang, and Yi Liu. Rt-detrv2: Improved baseline with bag-of-freebies for real-time detection transformer. *arXiv preprint arXiv:2407.17140*, 2024. 1

[4] Hanno Scharr, Massimo Minervini, Andreas Fischbach, and Sotirios A Tsaftaris. Annotated image datasets of rosette plants. In *European conference on computer vision*, pages 6–12. Suisse Zürich, 2014. 3, 4

[5] 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*, pages 3729–3738, 2023. 1

[6] Hideaki Uchiyama, Shunsuke Sakurai, Masashi Mishima, Daisaku Arita, Takashi Okayasu, Atsushi Shimada, and Rinichiro Taniguchi. An easy-to-setup 3d phenotyping platform for komatsuna dataset. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 2038–2045, 2017. 3, 4

[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 and machine intelligence*, 46(12):9583–9594, 2024. 3, 4

[8] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 2

[9] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2021. 1