

Negative-First: Self-Supervised Background Modeling for Plant Localization in Green-on-Green Grassland

Moritz Schauer¹ Daniel Dumschat¹ Ahmed Hammam² Seyed Eghbal Ghobadi¹

¹ University of Applied Sciences Mittelhessen, Germany · ² Aleph Alpha, Germany

AP_{PLANT}, BEST CONFIG
0.728

OF FULLY SUPERVISED AP
91%

MANUAL PLANT LABELS
0

END-TO-END, RTX 4090
51 FPS

01 The Problem: Green-on-Green

- Target plants hide in dense grasses, forage plants, soil and litter.
- Foreground and background share colour, texture and morphology — colour thresholds and vegetation indices fail.
- Boxes and masks are slow, expert work; a binary grass / not-grass call per cell is fast.



02 The Idea: Negative-First

Instead of labelling plants, annotate grass only — and treat everything that deviates from the grass distribution as a candidate plant.

Background material in grassland imagery is abundant and easy to verify locally, so modelling the negative class aligns supervision with the cheapest, most reliable data to collect. Frozen self-supervised backbones already separate structurally distinct regions without task-specific representation learning.

- No manually verified plant labels during training.
- Pseudo-positives mined by Mahalanobis distance to a grass-anchor background model.
- Localization as presence prediction on 64×64 px grid cells.
- Encoder frozen throughout; only a lightweight head is trained.

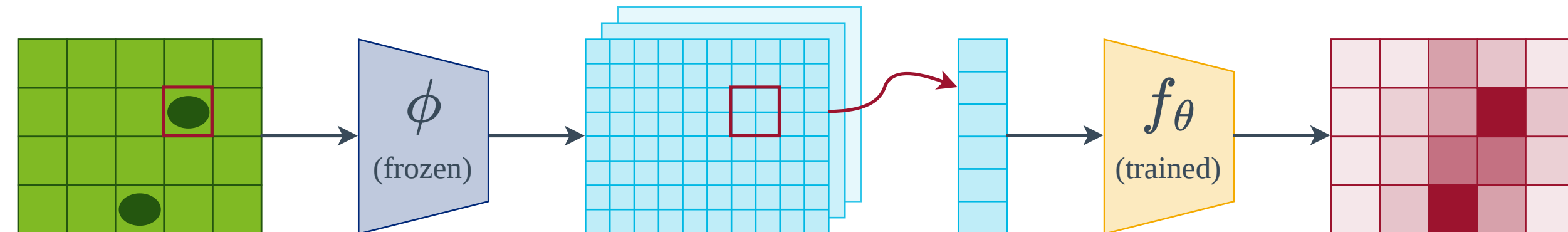


Fig. 1 Overview of the Negative-First pipeline. A grassland image x is encoded once by a frozen backbone Φ into a feature grid Z , finer than the label grid: each $s \times s$ cell g_{ij} covers a $k \times k$ block of local features ($k = 2$ for ConvNeXt-S at $s = 64$), pooled into a cell embedding z_{ij} and classified by head f_θ into presence probability p_{ij} ; stacking over the grid yields P . The highlighted cell traces one $g_{ij} \rightarrow z_{ij} \rightarrow p_{ij}$ path.

03 The Method

Precise boundaries are ill-posed in grassland: it is often unclear where a plant begins or ends. So we drop boundaries and ask one question per grid cell — is a target plant present? Cells are $s = 64$ px; at $0.5\text{--}0.9$ mm/px that is $3.2\text{--}5.8$ cm on the ground, fine enough to act on, coarse enough for robust binary labels.

A Anchor

Manual annotation covers grass-only windows; plant labels are never used in mining. These anchors are mean-pooled in frozen feature space.

B Model the background

The anchors define the grass model: a distribution over frozen features. Rescaling evens out the individual dimensions; a shrinkage covariance then captures the correlations between them, stable despite few anchors.

C Mine the tail

Artefact and artificial-colour windows are discarded. The rest are ranked by Mahalanobis distance from the grass model; the top 3% form the mining pool. Pseudo-positives are drawn from it at random, biased toward the most extreme — as many as there are grass anchors, so the set stays balanced.

D Train the head only

A small head maps the pooled cell descriptor to a plant-presence probability, and only this head is trained — on the mined set D_{NF} , or on the size-matched, fully labelled D_{FL} as a reference. Same pipeline both times; only the source of the positives differs.

Tail-enrichment assumption, not closed-world anomaly detection: not every outlier is a plant, but the high-score tail is enriched for plant evidence.

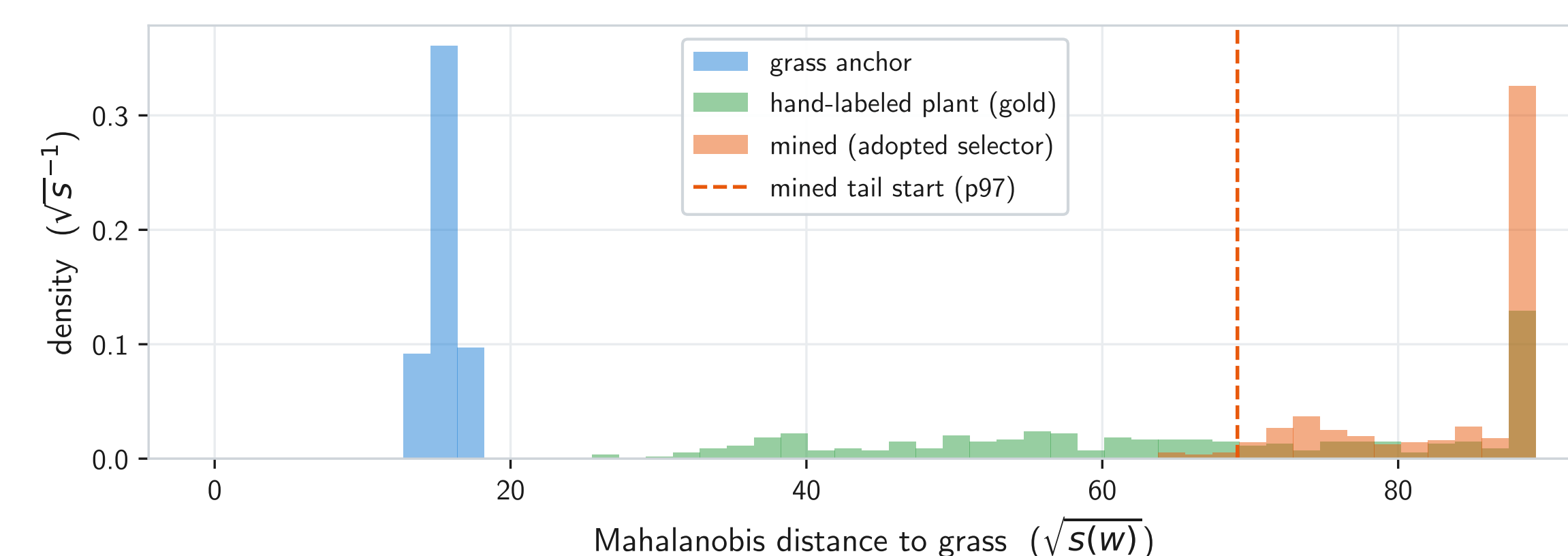


Fig. 2 Mined versus hand-labeled plants. Distribution of the Mahalanobis distance to the grass distribution for grass anchors, hand-labeled (gold) plants, and mined samples. Distances are computed in the frozen, mean-pooled DINOv3 ConvNeXt-S feature space using 64 px windows from the training frames on which mining is performed. The dashed line indicates the p_{97} tail threshold.

04 Results

Tab. 1 Dense consensus test set: 150 images, 81 000 manually labelled cells, 28.4% plant prevalence. Three seeds; mean pooling, evidential head.

Backbone	FULLY LABELED (D_{FL})		NEGATIVE-FIRST (D_{NF})	
	AP _{plant} ↑	AUROC _{plant} ↑	AP _{plant} ↑	AUROC _{plant} ↑
CNX-S	.804 ± .000	.903 ± .000	.728 ± .000	.869 ± .000
ViT-S	.707 ± .001	.850 ± .000	.652 ± .002	.831 ± .001
ImageNet CNX-S	.698 ± .002	.855 ± .001	.692 ± .001	.844 ± .001
2-means (label-free)	AP _{plant} 0.286			
Prevalence (predict-all)	AP _{plant} 0.284	AUROC _{plant} 0.500		



Fig. 3 Qualitative grid-level prediction. Predictions for a representative test frame using DINOv3 ConvNeXt-S at a visualization threshold of $\tau = 0.2$. Each cell corresponds to a 64×64 px window, and uncolored windows denote true negatives. The threshold is used solely for visualization and is not a tuned deployment operating point.

05 Takeaway

Deviation from a frozen-feature background model is a practical supervisory signal: frozen self-supervised representations encode enough morphological cue for label-efficient plant localization without a single plant annotation.

NEXT Adaptive background models, multi-scale localization, iterative pseudo-label refinement, uncertainty inside the backbone via LoRA, and evaluation across seasons, sensors and edge devices.



aimgrassland.mni.thm.de
moritz.schauer@mni.thm.de



Co-funded by
the European Union

