Unlocking Partially Annotated Agricultural Data: A Cut-and-Paste Data Augmentation Framework for Plant Detection

Michael Søndergaard Nørbo Madsen[1,2], René Gislum[2], and Rasmus Nyholm Jørgensen[1]

[1]Aarhus University, Department of Agroecology, Forsøgsvej 1, 4200 Slagelse, Denmark.[2]Aarhus University, Department of Electrical and Computer Engineering, Finlandsgade 22, 8200 Aarhus N, Denmark.

** Corresponding Author E-mail: msn@agro.au.dk*

A paper from the Proceedings of the

17th International Conference on Precision Agriculture and 11th Brazilian Congress on Precision and Digital Agriculture

13-15 July 2026

Porto Alegre, Brazil

Abstract

Accurate weed and crop recognition is essential for effective management practices in precision agriculture, enabling targeted herbicide application through automated spraying systems. However, the performance of deep learning models in real-world field settings is often limited by class imbalance, where broad categories such as monocotyledons and dicotyledons overshadow classes labelled at the species level. A major bottleneck in addressing this imbalance is the massive under-utilization of partially annotated data collected during operational field mapping. Using a large-scale commercial agricultural dataset as a practical test-case, we observed that while nearly 60,000 fully annotated images were utilized for training, over 70,000 additional images were excluded by standard object detection pipelines simply because they were only partially annotated. This rigid requirement left more than 330,000 expert-approved plant instances entirely unused.

To address this challenge, we developed a Thumbnail Injection Framework designed to leverage approved plant instances from partially annotated images. The framework systematically extracts these instances and incorporates them into images marked for training to enrich the dataset. To ensure realistic placement and maintain data integrity, the system utilizes an efficient grid-based representation of unoccupied space. A sliding-window search is employed to identify suitable injection locations that avoid overlapping with existing plant annotations or image boundaries. This allows for a more balanced representation of less frequent species without requiring any additional manual annotation.

While this framework was initially developed as an offline tool for dataset refinement, its capabilities have been extended to function as a dynamic training-time augmentation module. By integrating instance injection directly into the training pipeline, the object detection model is exposed to a more diverse variety of plant-on-soil combinations during each epoch. This approach specifically targets rare "wishlist" classes for detection: weed species that are critical for herbicide management but are typically under-represented in the training data.

The authors are solely responsible for the content of this paper, which is not a refereed publication. Citation of this work should state that it is from the Proceedings of the 17th International Conference on Precision Agriculture and 11th Brazilian Congress on Precision and Digital Agriculture. EXAMPLE: Last Name, A. B. & Coauthor, C. D. (2026). Title of paper. In Proceedings of the 17th International Conference on Precision Agriculture and 11th Brazilian Congress on Precision and Digital Agriculture (unpaginated, online). Monticello, IL: International Society of Precision Agriculture and Brazilian Congress on Precision and Digital Agriculture.

The framework was first established at operational scale in a prior large-scale study on Danish agricultural fields imaged at 5-24 MP (Nørbo Madsen, 2024). Compared with a control trained on the original dataset alone, offline injection of expert-approved thumbnails substantially improved detection in mAP@0.5:0.95 for rare classes, by 35.5% for creeping thistle (Cirsium arvense) and 48.8% for volunteer potato (Solanum tuberosum), with an overall gain of 17.9%. These results demonstrate that previously discarded data can be turned into useful training signal for site-specific weed control.

Building on this prior work, we extend the framework into a dynamic, training-time augmentation module and validate it in a new controlled single-region field study that isolates the mechanism. There, with a smaller detector and a single training seed, the dynamic variant produces consistent, dose-responsive gains on the targeted classes (a mean improvement of about 0.02-0.04 mAP@0.5 over five injected classes) without degrading overall performance, and it outperforms one-off offline injection. We further find that the fidelity and class balance of the recovered thumbnails are decisive to the effect, and that background masking helps or hurts depending on where a thumbnail was sourced.

Keywords.
copy-paste augmentation; object detection; computer vision; instance injection; partial annotation; class imbalance; weed detection; YOLO; precision agriculture; RoboWeedMaPS

Introduction

Weeds represent the single largest potential threat to global crop production, with estimated potential yield losses of up to 34%, exceeding those attributable to animal pests or pathogens (Oerke, 2006). The conventional response, field-wide broadcast application of herbicides, has come at a substantial ecological cost: decades of agricultural intensification and reliance on broadcast spraying have degraded functional in-field biodiversity and accelerated the evolution of herbicide-resistant weed populations. Together with increasingly ambitious regulatory frameworks, such as the European Union's Farm to Fork target of halving overall chemical pesticide use by 2030 (European Commission, 2020), these pressures have driven a shift toward precision Integrated Weed Management (IWM) and, in particular, site-specific weed control (SSWC), which transitions from uniform spraying to highly targeted, localized herbicide application (Christensen et al., 2009). The agronomic basis for this approach is well established: weed populations in arable fields are not distributed uniformly but occur in aggregated patches (Rew & Cousens, 2001; Wallinga et al., 1998), and these patches exhibit sufficient spatial and temporal stability across growing seasons that accurate mapping can substantially reduce total herbicide consumption (Gerhards et al., 1997; Johnson et al., 1996).

Recent advances in machine vision and robotic actuation have pushed the boundaries of SSWC from patch-level management down to the individual-plant level, and the agronomic viability of sub-decimeter, plant-by-plant spot spraying is no longer a theoretical concept but a proven commercial reality (Gerhards et al., 2022; Upadhyay et al., 2024). Evaluations of ultra-high-precision (UHP) smart sprayers, such as the EcoRobotix ARA system, demonstrate that targeted micro-dosing can drastically reduce chemical inputs by up to 90% relative to broadcast application, while simultaneously maximizing crop potential by reducing phytotoxicity (Anne et al., 2024).

However, the sheer eradication efficiency of these state-of-the-art UHP systems introduces a profound ecological dilemma. When field robotics are deployed solely to achieve absolute weed clearance, the remaining in-field biodiversity is critically threatened, and recent agroecological research emphasizes that weeding robots must be informed by ecological principles to reconcile the often-conflicting goals of securing crop productivity and protecting biodiversity (Zingsheim & Döring, 2024). Because arable weed communities are dynamic populations governed by recruitment, mortality, and frequency-dependent competition rather than static cover (Harper, 1967), a sustainable path forward requires exploiting within-field heterogeneity and species-specific traits, transitioning weed management from binary, total-field eradication toward a nuanced strategy of calculated coexistence. Realizing this strategy, deciding which species to remove and which to tolerate, hinges on accurate, timely recognition of weeds at the species level within the crop canopy.

Such species-level recognition has been made feasible by advances in image acquisition. High-resolution, sub-millimeter imaging is now possible in cereal crops at operational speeds of up to 50 km/h, as demonstrated by the platforms developed under the RoboWeedSupport and RoboWeedMaPS initiatives (Laursen et al., 2017). Such imagery enables the extraction of the fine-grained shape and texture characteristics needed to discriminate species (Gerhards et al., 1998). When coupled with dedicated computer-vision algorithms, for example, the monocot and dicot coverage ratio vision (MoDiCoVi) algorithm, these systems have achieved herbicide reductions of up to 65% in maize without measurable loss of biological efficacy (Laursen et al., 2016). State-of-the-art SSWC increasingly relies on deep learning, and in particular on single-stage object detectors of the YOLO family, which simultaneously localize and classify multiple plant instances in real time (Redmon et al., 2016).

The performance of these models, however, is fundamentally constrained by the data on which they are trained. Robust deep neural networks require large, accurately annotated datasets, yet manual annotation of field imagery is labor-intensive, error-prone, and subject to substantial intra- and inter-annotator variability (Dyrmann et al., 2016). Annotation at the species level is typically organized using standardized plant codes such as the EPPO coding system, which imposes a natural hierarchy on the labels. In operational datasets, this hierarchy produces a severe class imbalance: broad parent categories such as "monocotyledons" and "dicotyledons" overwhelmingly dominate, overshadowing the specific, lower-frequency species classes that matter most for targeted chemical intervention. This foreground-foreground imbalance is a well-recognized obstacle to generalization in object detection (Oksuz et al., 2021; Johnson & Khoshgoftaar, 2019), and in the SSWC context it directly degrades the model's ability to confidently identify the rare weed species that drive spraying decisions.

A further bottleneck compounds this imbalance. Standard object detection pipelines generally demand fully annotated images, yet field-collected images frequently contain a mixture of expert-approved annotations alongside unapproved or uncertain labels. Rigid pipeline constraints force the exclusion of these partially annotated images, causing massive under-utilization of operational data. In the large-scale commercial dataset used as the practical test-case for this work, nearly 60,000 fully annotated images were available for training while more than 70,000 partially annotated images were discarded, leaving over 330,000 expert-approved plant instances entirely unused.

Several strategies have been proposed to extract value from imperfectly labeled agricultural data, including semi-supervised learning and pseudo-labeling (Li et al., 2024) and the generation of synthetic training imagery (Sapkota et al., 2022; Hasan et al., 2023). A particularly effective and architecturally simple line of work in computer vision composes new training images by cutting object instances and pasting them onto other images; such "copy-paste" augmentation has proven to be a strong, low-cost method that is especially beneficial for rare categories (Dwibedi et al., 2017; Ghiasi et al., 2021). To date, however, these techniques have not been adapted to recover expert-approved instances from the partially annotated images that operational mapping pipelines routinely reject, precisely the data reservoir that is both abundant and freely available.

To address this gap, this paper introduces the Thumbnail Injection Framework, a domain-specific cut-and-paste data augmentation approach for plant detection. The framework systematically extracts expert-approved plant instances ("thumbnails") from unused, partially annotated images and incorporates them into images marked for training. To ensure realistic placement and preserve data integrity, it represents unoccupied image space with an efficient grid and employs a sliding-window search to identify injection locations that avoid overlap with existing annotations or image boundaries, enabling a more balanced representation of infrequent species without any additional manual annotation. While the framework was first developed as an offline tool for dataset refinement (Nørbo Madsen, 2024), we extend it here into a dynamic, training-time augmentation module that exposes the detector to a more diverse range of plant-on-soil combinations at every epoch and specifically targets underrepresented "wishlist" species critical for herbicide management. The offline tool was first validated at operational scale, with substantial gains in mAP@0.5:0.95 for rare classes including a 35.5% improvement for creeping thistle (*Cirsium arvense*) and a 48.8% improvement for volunteer potato (*Solanum tuberosum*) relative to a control model trained on the original dataset alone (Nørbo Madsen, 2024); here we test whether the same mechanism, delivered dynamically, holds up in a controlled study designed to isolate it.

Methods

The Thumbnail Injection Framework

The framework turns approved instances that the pipeline would otherwise throw away into extra training signal for the rare classes. Here, *rare* means the low-frequency, species-level classes (for example thistle or volunteer potato), as opposed to the dominant, broad parent classes such as monocotyledons and dicotyledons.

It has three parts. A pool supplies the raw material: approved thumbnails harvested from the images that training rejects. A placement step then decides where each thumbnail goes, leaving the labels already in the image untouched. Two delivery modes apply the thumbnails, either offline by baking them into a static training set, or on the fly inside the training loop.

Operational pipelines train only on fully annotated images (all plant instances labeled with class and bounding box) and discard the rest, even when a rejected image carries perfectly good, expert-approved boxes. Even some fully annotated images carry plant instances approved only at a higher hierarchy level, such as monocotyledon or dicotyledon. We treat the rejected images as a donor pool. For each target class, every approved box on a rejected image becomes a thumbnail, indexed by its EPPO code. Nothing new is asked of an annotator; only already-approved instances are used, and the unverified parts of a donor image never reach the detector.

A thumbnail has to land somewhere free of existing boxes and inside the frame. To find such a spot quickly, the free area of each training image is mapped onto a coarse grid, and any cell covered by an existing box, or by a thumbnail already placed, is marked occupied. A window the size of the candidate thumbnail then slides across the free cells from a random start, and the thumbnail is dropped at the first position where the whole window is clear. Working on a grid whose cells are *n* pixels on a side, rather than pixel by pixel, reduces the search from order *W*H* to:

$$cost \;\approx\; (W * H) \,/\, n^2 \tag{1}$$

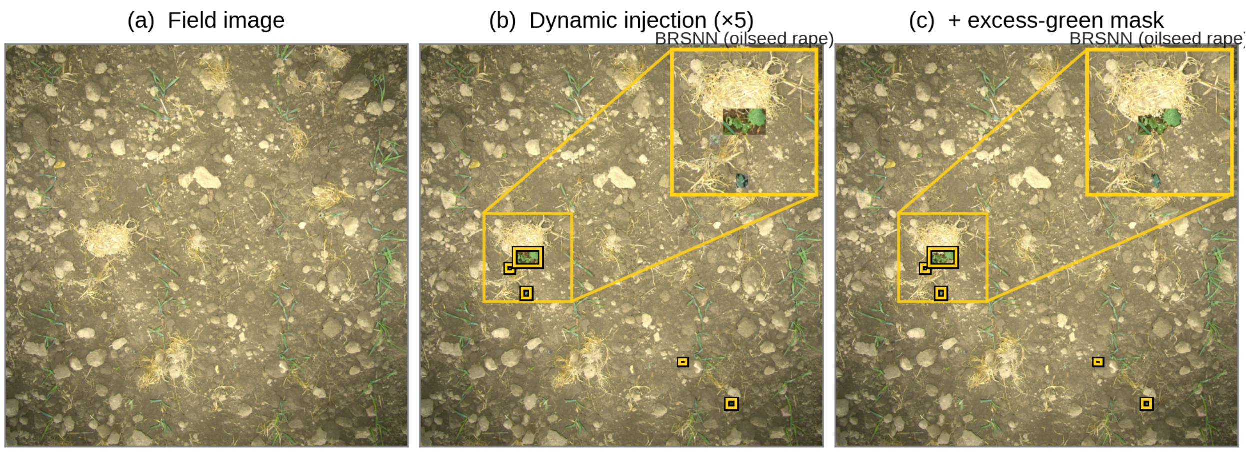
for an image of width *W* and height *H*. Each placed thumbnail adds one box of its own class, and nothing else in the label file changes. Field backgrounds are largely uniform, so geometric placement is enough on its own and no learned model of visual context is needed (Dvornik et al., 2018).

In offline mode the framework writes a single enriched copy of the training set: the thumbnails are injected once, and training then runs normally on the augmented data. This is the dataset-refinement tool of the original framework (Nørbo Madsen, 2024).

In dynamic mode the injection moves inside the training loop and runs as a callback while each image loads. With probability *p_inj*, a few thumbnails are sampled, prepared, placed and pasted, and their boxes are appended before the image reaches the detector's own augmentation stage (mosaic, flips). Because the sampling is repeated every epoch, the detector meets a fresh set of plant-on-soil combinations on each pass instead of a fixed, pre-baked one. The per-class priors stay fixed (uniform over the targets), and the module reads nothing back from the model; it is augmentation, not a controller. One detail matters for reproduction: the data-loader's worker processes have to be re-initialised after the callback is registered, or the injection runs in the main process alone and is silently skipped in the workers.

Three controls govern thumbnail quality. The first is scale. Thumbnails and base images are resampled to a common ground sample distance (millimetres per pixel), so that a pasted plant appears at its true physical size rather than at an arbitrary pixel size; train-test resolution mismatch is a known source of degradation in agricultural detection (Velumani et al., 2021). The second is sharpness: a thumbnail may shrink but is never blown up past its native size, which keeps it crisp instead of blurred. The third is balance, with the pool holding equal numbers of thumbnails per class so that an abundant class cannot swamp the rest. On top of these, each paste receives light augmentation: small scale jitter, horizontal and vertical flips, and mild brightness and contrast changes, but no rotation, which would expose the corners of the cut-out patch. An optional excess-green mask can retain the plant pixels and drop the surrounding soil (Woebbecke et al., 1995; Meyer & Neto, 2008); we test it as an ablation (Figure 1). The controlled experiment used all three controls; the operational-scale experiment pasted at native scale without them.

Figure 1. Dynamic injection of a recovered thumbnail onto a training image, without (b) and with (c) excess-green background masking.



Datasets

All imagery comes from RoboWeedMaPS (RWM), a commercial collection of top-down Danish field images (5-24 MP) with expert, EPPO-coded bounding boxes. EPPO codes run from broad parents (undifferentiated monocotyledons and dicotyledons) down to individual species, and in operational data the broad parents dominate. The species classes that actually drive a spraying decision sit at the thin end of that distribution, and they are what the framework targets. The RWM collection and its predecessors have underpinned a range of weed-assessment and integrated-weed-management studies, from annotator-variability analysis (Dyrmann et al., 2016) and the Open Plant Phenotype Database (Leminen Madsen et al., 2020) to weed mapping matched to sprayer capacity (Somerville et al., 2019), the image counts needed for robust spray maps (Somerville et al., 2021), and monocot/dicot detection in dense field environments (Teimouri et al., 2022).

The waste is large. Of the operational training pool, nearly 60,000 fully annotated images were used while more than 70,000 partially annotated ones were set aside, resulting in over 330,000 approved instances not being used for training.

At operational scale (Experiment 1), the detector was trained on the full approved-for-training pool (on the order of 60,000 fully annotated images drawn from many fields and seasons), with a separate held-out test split. Five rare targets were injected:

- CIRAR (creeping thistle),
- SPQOL (spinach),
- VICFX (field bean),
- SOLTU (volunteer potato)
- PIBSA (field pea)

using 35,524 thumbnails taken from the excluded, partially annotated images.

The controlled experiment (Experiment 2) used all approved-for-training field images from a single region, AU-Flakkebjerg. To keep the evaluation clean, controlled-condition phenotyping imagery and a set of flagged uploads were removed from this base, leaving 6,390 images across 10 classes, resampled to a common 0.35 mm/px and split upload-disjoint into roughly 70/15/15 (4,753 / 822 / 815 images). Five scarce classes were injected: CHEAL (fat-hen), CIRAR (creeping thistle), VERPE (field speedwell), BRSNN (volunteer oilseed rape), and POLAV (knotgrass); the remaining six were left as controls. The thumbnail pool was balanced to 2,500 sharp instances per class and drawn only from approved boxes on uploads that share no image with the train, validation or test splits, so that no pasted instance can leak into evaluation. Because several targets are scarce in Flakkebjerg itself, this pool spans field uploads from across the wider RWM collection and, for the two scarcest classes, the controlled phenotyping imagery that the base excludes (Leminen Madsen et al., 2020): VERPE is drawn predominantly from that controlled imagery (about 80% of its pool), POLAV only partly (about 18%), while CHEAL, CIRAR and BRSNN are sourced entirely from field uploads. The injected instances therefore come, by construction, from different fields, cameras, seasons and conditions than the Flakkebjerg images they are pasted into and tested against. A second property of the data matters for interpretation: RWM labels are applied at inconsistent levels of the EPPO hierarchy, so a dicot plant may be annotated at species level (for example CHEAL) or left at the broad parent class PPPDD, with the two coexisting as mutually exclusive detector classes.

Of the five injected targets, four (fat-hen, field speedwell, volunteer oilseed rape and knotgrass) are dicot weeds that an annotator may also leave under the generic parent PPPDD; creeping thistle (CIRAR), the most distinctive of the five, is in practice the one most reliably annotated at species level. As Table 1 shows, the injected classes are scarce in training and scarcer still in test (CIRAR, for instance, has 746 training and only 91 test instances), which limits how finely their effect can be read.

Table 1. Controlled-dataset (AU-Flakkebjerg) class composition. Injected classes marked with an asterisk (*). The remaining six are controls. PPPDD and PPPMM are the broad dicot and monocot parent classes.

Class	Train	Val	Test
PPPDD (dicot)	208,796	45,764	53,613
PPPMM (monocot)	15,320	3,292	4,456
FAGES	13,347	3,559	4,285
CHEAL*	6,592	2,267	2,393
SOLTU	3,095	131	271
ZEAMX	955	182	381
BRSNN*	953	233	243
CIRAR*	746	134	91
VERPE*	628	148	113

POLAV*	618	127	158

Experimental design and evaluation

Experiment 1 trained a YOLOv5l detector (Ultralytics, 2020) at 1600 px for 300 epochs, and compared a control model against the same model trained on data enriched offline with injected thumbnails, in a single run. This is the configuration of (Nørbo Madsen, 2024): a large architecture, a long schedule, the full operational dataset.

Experiment 2 trained a smaller YOLOv11m detector (Ultralytics, 2024) at the same 1600 px, but for 100 epochs and on a single seed (42). We turned off the detector's own copy-paste and mixup, so that injected thumbnails were the only pasted instances and any effect could be traced to them; mosaic stayed on. Four arms were compared: a no-injection baseline, offline injection, dynamic injection at three doses (one, three or five thumbnails per image, written x1, x3, x5, as far as free space allowed), and dynamic x5 with excess-green masking. The sampling priors were uniform over the five targets, with $p_{inj} = 1.0$. Two pools were used throughout, a naive pool (imbalanced, upscaled) and the quality-controlled pool (balanced, sharp), which isolates what thumbnail quality alone contributes.

Every detector was scored on the clean test split, which never receives injected instances. We report per-class average precision at IoU 0.5 (AP@0.5) and the COCO-style average over 0.5-0.95 (mAP@0.5:0.95), the mean over all classes, and the mean over the five injected classes as a summary of the targeted effect. With a single seed, we read direction, dose-response and consistency across classes rather than significance.

Results

We ran two evaluations that answer two different questions. The first (Experiment 1) is our prior operational-scale study (Nørbo Madsen, 2024), re-reported here as deployment context; it asks whether the recovered data pays off in a realistic deployment. The second (Experiment 2) is this paper's new contribution: it asks whether the effect is real and whether the new dynamic delivery earns its place, and it was kept deliberately small: a smaller detector (YOLOv11m, ~20M parameters) and 100 training epochs. This is a third of the operational schedule, and short enough that the dynamic runs had not finished improving; it uses a single region, rare-class test sets as small as 91 instances, and one seed. Those choices cut both the size of any gain and our ability to measure it, so the controlled numbers below are best read as a floor, not a ceiling.

Experiment 1: Operational scale

At the operational scale, offline injection helped four of the five targeted classes (Table 2). The gains were largest where the class was rarest: creeping thistle rose 35.5% and volunteer potato 48.8% in mAP@0.5:0.95, and the overall figure improved 17.9%. One class, PIBSA, went the other way and fell 15.3%. All of this came from a large detector trained to convergence on the operational data, in a single run.

Table 2. Operational-scale offline injection, control → injected (Nørbo Madsen, 2024).

Class (eppo)	Thumbnails	mAP@0.5 (Δ%)	mAP@0.5:0.95 (Δ%)
CIRAR	+14,009	0.779 → 0.900 (+15.5%)	0.581 → 0.787 (+35.5%)
SPQOL	+11,678	0.841 → 0.900 (+7.0%)	0.607 → 0.753 (+24.0%)
VICFX	+7,892	0.892 → 0.914 (+2.5%)	0.693 → 0.778 (+12.3%)

SOLTU	+1,117	0.383 → 0.518 (+35.2%)	0.256 → 0.381 (+48.8%)
PIBSA	+828	0.872 → 0.784 (−10.1%)	0.575 → 0.487 (−15.3%)
Overall		0.785 → 0.838 (+6.8%)	0.557 → 0.657 (+17.9%)

Experiment 2: Controlled evaluation

On the quality-controlled pool, dynamic injection lifted the injected-class mean by 0.023, 0.033 and 0.024 AP@0.5 at x1, x3 and x5, with the moderate x3 dose best, and it nudged the overall 10-class score from flat to slightly positive (+0.013 at x3). The targeted classes improved, in other words, without costing the rest of the detector anything. CHEAL is the one injected class with a large test set (n = 2,393), and so the one we trust most; its AP@0.5 climbed from 0.376 to 0.452, about a fifth higher. CIRAR rose from 0.204 to 0.244.

The mAP@0.5:0.95 picture is the same, with the injected-class mean up 0.024 at x3. Four of the five injected classes improved. BRSNN was the exception, and gained nothing at any dose (Table 3).

Table 3. Controlled dynamic injection on the quality-controlled pool, AP@0.5; n = 1. Δ rows are versus the no-injection baseline.

Class	Test *n*	Baseline	x1	x3	x5	x5 + mask
CHEAL	2,393	0.376	0.419	0.413	0.452	0.394
CIRAR	91	0.204	0.229	0.244	0.237	0.206
VERPE	113	0.172	0.191	0.165	0.189	0.196
BRSNN	243	0.163	0.103	0.169	0.117	0.127
POLAV	158	0.293	0.385	0.382	0.332	0.382
Injected mean	-	0.242	0.265	0.274	0.265	0.261
Overall	-	0.326	0.331	0.340	0.338	0.324
Δ injected mean	-	-	+0.023	+0.033	+0.024	+0.019
Δ overall	-	-	+0.004	+0.013	+0.012	-0.002

Delivery mode mattered. Holding the pool fixed at the earlier naive pool (the only iteration that also carries an offline arm), offline injection actually lowered both the injected-class mean (−0.005) and the overall score (−0.015), while dynamic injection raised the injected-class mean by as much as 0.031 and left the overall score intact (Table 4). The benefit comes from re-drawing fresh combinations every epoch, not from the extra instances on their own. That is the case for moving injection into the training loop rather than baking it in once. Table 3 isolates *dose* with the good pool; Table 4 isolates *delivery mode* on the matched naïve pool - both share the same no-injection baseline (injected mean 0.242, overall 0.326).

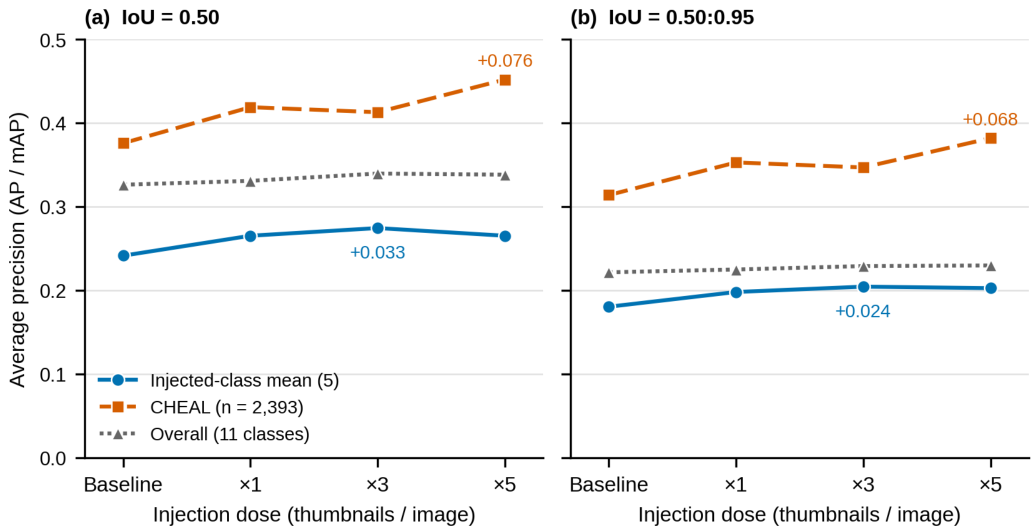
Table 4. Offline vs dynamic delivery on a single matched pool (the earlier naïve pool, the only iteration with an offline arm), AP@0.5; n = 1.

Arm	Injected mean (Δ)	Overall (Δ)
Baseline	0.242	0.326
Offline	0.237 (-0.005)	0.312 (-0.015)
Dynamic x1	0.255 (+0.013)	0.324 (-0.003)
Dynamic x3	0.252 (+0.010)	0.318 (-0.008)
Dynamic x5	0.273 (+0.031)	0.327 (+0.001)

Thumbnail quality turned out to be decisive. With the naive pool, which was imbalanced and upscaled until it blurred, dynamic injection left CIRAR flat or worse (−0.021, −0.083 and −0.003 across the three doses). Swapping in the sharp, balanced pool pulled CIRAR back to +0.025, +0.040 and +0.033, raised the injected-class mean, and turned the overall effect from slightly negative to non-negative. Sharpness and balance are not finishing touches here, they are what makes the method work at all.

Background masking cut both ways, depending on where a thumbnail came from and on the species itself. Masking helped two classes by removing soil without cost: VERPE (+0.007), whose pool is largely controlled phenotyping imagery with a uniform background worth removing, and POLAV (+0.050). On the two field-sourced classes it hurt: CHEAL (−0.059) and CIRAR (−0.031). The lesson is to mask selectively rather than across the board; the Discussion examines why.

Figure 2. Dose-response of dynamic injection on the quality-controlled pool: injected-class mean, CHEAL, and the overall against dose, at (a) IoU 0.50 and (b) IoU 0.50:0.95.



(a) IoU = 0.50
(b) IoU = 0.50:0.95
Average precision (AP / mAP)
0.0
0.1
0.2
0.3
0.4
0.5
+0.076
+0.033
+0.068
+0.024
Injected-class mean (5)
CHEAL (n = 2,393)
Overall (11 classes)
Baseline
×1
×3
×5
Injection dose (thumbnails / image)

Discussion

Cutting object instances out of one image and pasting them into another is an old and well-understood augmentation, and it is known to help rare classes (Dwibedi et al., 2017; Ghiasi et al., 2021). Our contribution is not the paste. It is where the instances come from and how they are delivered. We pull them out of the partially annotated images that the pipeline throws away, and we inject them dynamically, during training, into an agricultural detector. The nearest in-domain work pastes weed pixels on a YOLO detector but sources them from fully labelled images (Wang et al., 2022); the nearest dynamic copy-paste places instances under model guidance and is built for aerial rather than field imagery (Li et al., 2025).

Our setting draws on both while matching neither. Like the in-domain work it operates on top-down field imagery and an agricultural detector, but it sources its pastes from the partially annotated images the pipeline discards rather than from fully labelled ones; like the dynamic work it re-samples instances every epoch, but it delivers them through a fixed-prior augmentation module rather than a model-guided controller, and on field rather than aerial scenes. Recovering an approved paste source from partial annotations is the element that is genuinely new, and it is also the one that does the work, converting an otherwise discarded reservoir of expert-approved instances into training signal at no additional labelling cost.

The two experiments sit at opposite ends of a resource scale. The operational run (prior work; Nørbo Madsen, 2024) had a large detector, a 300-epoch schedule and the full, geographically varied dataset behind it, and it produced large gains. The controlled run had a smaller model, a third of the training, a single region, and rare-class test sets one to two orders of magnitude smaller, conditions that should suppress an effect, not invent one. A consistent, dose-responsive signal still came through: dynamic beating offline, CIRAR recovered, the rest of the detector untouched in aggregate. That the signal survives such an unfavourable setting is what matters. It reads as a lower bound on a real effect, and the large operational gains look like the same mechanism without the constraints. Because the two settings differ in model, budget and scope, we compare them by direction and consistency, not by absolute number.

Two properties of the controlled setting work against detecting the injection signal, and both imply the measured effect is conservative. The first is label-hierarchy noise. The training data hold more than 200,000 dicot instances under a single broad PPPDD class, and that class is itself a mixed bag: in earlier work on RWM data, a precision-controlled classifier could confidently re-assign 26.7% of generic PPPDD (dicot) thumbnails to a specific species (Nørbo Madsen, 2024). Because PPPDD and the species classes are trained as mutually exclusive, the detector is penalised for confidently calling a plant CHEAL whenever an equivalent plant elsewhere carries only the parent label, which caps how confidently the parent-ambiguous species can be scored and adds gradient noise regardless of injection. Four of the five injected targets are dicot weeds exposed to this competition; CIRAR, the most distinctive, is the one most reliably annotated at species level rather than absorbed into the parent. Injecting more instances of a class cannot fully overcome a label conflict of this kind, so the conflict compresses and obscures the signal we are trying to read. This competition is also weaker in the operational study, where four of the five targets were crops (spinach, faba bean, potato, field pea) imaged and labelled at species level in crop contexts and thus rarely left as generic dicot, one reason that study's gains read cleaner. We do not over-read the per-class pattern, however: CHEAL is parent-ambiguous yet improves the most, so the hierarchy noise suppresses the effect rather than erasing it.

The second property is a domain gap between the injected instances and the evaluation. To balance the scarce classes, the thumbnail pool is drawn from uploads outside Flakkebjerg, and for VERPE largely (and POLAV partly) from controlled phenotyping imagery, so the pasted instances differ in field, camera, resolution and lighting from the single-region images they are added to and tested on. At operational scale, where the detector already sees tens of thousands of images from many fields and cameras, this diversity is an asset and is the likely source of the large offline gains; at the controlled scale, with few instances injected and a single-region test set, a handful of visually divergent but correctly labelled instances can act as much as distractors as useful positives. The evaluation compounds this, because a single-region test set cannot reward the cross-domain generalisation that off-domain instances are best placed to provide.

The background-masking result follows from the same logic, through two mechanisms that point the same way. The first is missing context: stripping the soil also strips the field context a real-field paste relies on, so masking hurts thumbnails cut from field images. The second follows directly from how the mask is defined. Excess-green masking thresholds on green chromaticity and so excludes achromatic and low-green pixels by construction (Woebbecke et al., 1995; Meyer & Neto, 2008). For species whose discriminative appearance resides partly in such pixels (the glaucous, farinose bloom of C. album, its anthocyanin-tinted stems, and the pale dentate margins of *C. arvense*), the mask removes part of the plant's own diagnostic signal along with the soil. Both mechanisms act on the two field-sourced, partly non-green targets (CHEAL, CIRAR) and explain why masking degraded them. The two classes it helped are uniformly green: VERPE is sourced from clean controlled-background imagery, so the threshold removes background only; POLAV, although mostly field-sourced, carries no non-green diagnostic feature to lose and apparently depends less on soil context, so soil removal helps it despite its field origin. The prescription, then, is to mask selectively, applying it only when a thumbnail's background is uniform and the species' diagnostic appearance is fully green, rather than across the board. Read together, label-hierarchy competition and the injected-instance domain gap both bias the controlled result downward, which is consistent with treating it as a floor and with the much larger gains the same framework gives at operational scale.

Three findings carry straight into practice:

1) Deliver the injection dynamically rather than baking it in once.
2) Keep the thumbnails sharp and the pool balanced, because a careless pool can erase the benefit entirely.
3) Mask backgrounds only when the thumbnails come from a controlled, uniform-background source and the species is uniformly green; for field thumbnails, leave the soil where it is.

A moderate dose - *around three thumbnails an image* - was the most dependable.

Several limits qualify all of this. The controlled results rest on one seed, so we report direction and consistency rather than significance, and proper error bars wait on multi-seed runs. Only CHEAL has a test set large enough to pin down; the other injected classes are directional. BRSNN never benefits at any dose, which the two effects above may bear on but do not fully explain. Free-space placement caps the dose, so the highest nominal setting injects fewer instances than its label suggests. In aggregate the controlled injection left the detector intact, but it did incidentally depress a few un-injected classes (notably SOLTU and ZEAMX) through paste clutter, offset by gains on the abundant classes. Pasting without blending leaves a further, residual gap, and the quality of a recovered thumbnail is not guaranteed: in the operational study the one regression, PIBSA, was traced to its cut-outs. Field pea grows intertwined with neighbouring plants, so a PIBSA bounding box frequently encloses fragments of other species, and injecting such mixed cut-outs feeds the detector label-noise rather than clean positives (Nørbo Madsen, 2024). PIBSA also entered that experiment already well represented (over 10,000 training instances), so its ~800 injected thumbnails added little but their noise, a concrete reminder that recovered instances inherit the imperfections of real, operational annotation. Finally, the two experiments use different detectors and different data, and are not comparable head to head.

The obvious next step is to run the dynamic module at full operational scale, with a larger architecture and a longer schedule, where the offline results suggest the gains should be largest. Beyond that lie several directions: multiple seeds for significance; removing the label-hierarchy conflict, either with a hierarchy-aware or partial-label loss or by first re-labelling the PPPDD class to species, so the injection signal is not masked by parent and child competing; drawing the pool from the same domain as the base where enough instances exist, or closing the domain gap with light blending or domain adaptation; source-aware masking and other realism work; priors that favour the species a grower most wants to control; and a closed-loop version that adjusts the injection rate to the detector's current weak spots.

Conclusion

Partially annotated field imagery is not empty. It is full of expert-approved instances that the standard pipeline simply never reaches. Recover those instances, paste them back into training, and the rare classes improve at no extra labelling cost. In a small, controlled setting (a compact model, a short schedule, one region and one seed), the new dynamic variant still produced a steady gain on the targeted classes and still beat offline delivery, without dragging down the detector overall. Set beside the large gains the same framework gives offline at operational scale (Nørbo Madsen, 2024), the message is consistent: this is a cheap, annotation-free way to detect the rare species that matter most, and it should only grow stronger with bigger models, longer training and more data. For the very common case of a detection dataset full of partially annotated images, that is a practical lever worth pulling.

Acknowledgements

The authors acknowledge the RoboWeedMaPS service and dataset for providing the agricultural field imagery and expert annotations central to this work. The framework, baseline experiments and methodology are documented in the first author's MSc thesis, openly archived at https://doi.org/10.5281/zenodo.18801981 (Nørbo Madsen, 2024). This work was supported by the Novo Nordisk Foundation (grant NNF23SA0086122, System-based Precision Agriculture for Sustainable Crop Production).

References

Anne, P., Gasser, S., Göttl, M., & Tanner, S. (2024). The reduction of chemical inputs by ultra-precise smart spot sprayer technology maximizes crop potential by lowering phytotoxicity. *Frontiers in Environmental Economics, 3*, 1394315. https://doi.org/10.3389/frevc.2024.1394315

Christensen, S., Søgaard, H. T., Kudsk, P., Nørremark, M., Lund, I., Nadimi, E. S., & Jørgensen, R. (2009). Site-specific weed control technologies. *Weed Research, 49* (3), 233-241. https://doi.org/10.1111/j.1365-3180.2009.00696.x

Dvornik, N., Mairal, J., & Schmid, C. (2018). Modeling visual context is key to augmenting object detection datasets. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 375-391). https://doi.org/10.1007/978-3-030-01258-8_23

Dwibedi, D., Misra, I., & Hebert, M. (2017). Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 1310-1319). https://doi.org/10.1109/ICCV.2017.146

Dyrmann, M., Midtiby, H. S., & Jørgensen, R. N. (2016). Evaluation of intra variability between annotators of weed species in color images. Paper presented at the International Conference on Agricultural Engineering (CIGR-AgEng 2016), Aarhus, Denmark, 26-29 June 2016.

European Commission. (2020). *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: A Farm to Fork Strategy for a fair, healthy and environmentally-friendly food system* (COM(2020) 381 final). Brussels: European Commission.

Gerhards, R., Wyse-Pester, D. Y., Mortensen, D. A., & Johnson, G. A. (1997). Characterizing spatial stability of weed populations using interpolated maps. *Weed Science, 45*(1), 108-119. https://doi.org/10.1017/S0043174500092559

Gerhards, R., Juroszek, P., Klümper, H., & Kühbauch, W. (1998). Möglichkeiten zur photobiologischen Unkrautregulierung auf Ackerschlägen. *Pflanzenbauwissenschaften, 2*(2), 91-96.

Gerhards, R., Andújar Sanchez, D., Hamouz, P., Peteinatos, G. G., Christensen, S., & Fernandez-Quintanilla, C. (2022). Advances in site-specific weed management in agriculture: A review. *Weed Research, 62*(2), 123-133. https://doi.org/10.1111/wre.12526

Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.-Y., Cubuk, E. D., Le, Q. V., & Zoph, B. (2021). Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2918-2928). https://doi.org/10.1109/CVPR46437.2021.00294

Harper, J. L. (1967). A Darwinian approach to plant ecology. *The Journal of Ecology, 55*(2), 247-270. https://doi.org/10.2307/2257876

Hasan, A. S. M. M., Diepeveen, D., Laga, H., Jones, M. G. K., & Sohel, F. (2023). Image patch-based deep learning approach for crop and weed recognition. *Ecological Informatics, 78*, 102361. https://doi.org/10.1016/j.ecoinf.2023.102361

Johnson, G. A., Mortensen, D. A., & Gotway, C. A. (1996). Spatial and temporal analysis of weed seedling populations using geostatistics. *Weed Science, 44*(3), 704-710. https://doi.org/10.1017/S0043174500094571

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data, 6*, 27. https://doi.org/10.1186/s40537-019-0192-5

Laursen, M. S., Jørgensen, R. N., Midtiby, H. S., Jensen, K., Christiansen, M. P., Giselsson, T. M., Mortensen, A. K., & Jensen, P. K. (2016). Dicotyledon weed quantification algorithm for selective herbicide application in maize crops. *Sensors, 16*(11), 1848. https://doi.org/10.3390/s16111848

Laursen, M. S., Jørgensen, R. N., Dyrmann, M., & Poulsen, R. (2017). RoboWeedSupport - Sub millimeter weed image acquisition in cereal crops with speeds up till 50 km/h. *International Journal of Biological, Biomolecular, Agricultural, Food and Biotechnological Engineering, 11*(4), 317-321. https://doi.org/10.5281/zenodo.1130297

Leminen Madsen, S., Mathiassen, S. K., Dyrmann, M., Laursen, M. S., Paz, L.-C., & Jørgensen, R. N. (2020). Open Plant Phenotype Database of common weeds in Denmark. *Remote Sensing, 12*(8), 1246. https://doi.org/10.3390/rs12081246

Li, J., Chen, D., Yin, X., & Li, Z. (2024). Performance evaluation of semi-supervised learning frameworks for multi-class weed detection. *Frontiers in Plant Science, 15*, 1396568. https://doi.org/10.3389/fpls.2024.1396568

Li, Z., Lian, S., Pan, D., Wang, Y., & Liu, W. (2025). AD-Det: Boosting object detection in UAV images with focused small objects and balanced tail classes. *Remote Sensing, 17*(9), 1556. https://doi.org/10.3390/rs17091556

Meyer, G. E., & Neto, J. C. (2008). Verification of color vegetation indices for automated crop imaging applications. *Computers and Electronics in Agriculture, 63*(2), 282-293. https://doi.org/10.1016/j.compag.2008.03.009

Nørbo Madsen, M. S. (2024). *Leveraging hierarchical classification and thumbnail injection for improved weed and crop detection* [MSc thesis, Aarhus University]. Zenodo. https://doi.org/10.5281/zenodo.18801981

Oerke, E.-C. (2006). Crop losses to pests. *The Journal of Agricultural Science, 144*(1), 31-43. https://doi.org/10.1017/S0021859605005708

Oksuz, K., Cam, B. C., Kalkan, S., & Akbas, E. (2021). Imbalance problems in object detection: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(10), 3388-3415. https://doi.org/10.1109/TPAMI.2020.2981890

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 779-788). https://doi.org/10.1109/CVPR.2016.91

Rew, L. J., & Cousens, R. D. (2001). Spatial distribution of weeds in arable crops: Are current sampling and analytical methods appropriate? *Weed Research, 41*(1), 1-18. https://doi.org/10.1046/j.1365-3180.2001.00215.x

Sapkota, B. B., Popescu, S., Rajan, N., Leon, R. G., Reberg-Horton, C., Mirsky, S., & Bagavathiannan, M. V. (2022). Use of synthetic images for training a deep learning model for weed detection and biomass estimation in cotton. *Scientific Reports, 12*, 19580. https://doi.org/10.1038/s41598-022-23399-z

Somerville, G. J., Jørgensen, R. N., Bojer, O. M., Rydahl, P., Dyrmann, M., Andersen, P., Jensen, N.-P., & Green, O. (2019). Marrying futuristic weed mapping with current herbicide sprayer capacities. In *Precision Agriculture '19* (pp. 231-237). Wageningen Academic Publishers. https://doi.org/10.3920/978-90-8686-888-9_28

Somerville, G. J., Mathiassen, S. K., Melander, B., Bøjer, O. M., & Jørgensen, R. N. (2021). Analysing the number of images needed to create robust variable spray maps. *Precision Agriculture, 22*(5), 1377-1396. https://doi.org/10.1007/s11119-021-09800-3

Teimouri, N., Jørgensen, R. N., & Green, O. (2022). Novel assessment of region-based CNNs for detecting monocot/dicot weeds in dense field environments. *Agronomy, 12*(5), 1167. https://doi.org/10.3390/agronomy12051167

Ultralytics. (2020). *YOLOv5* [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.3908559

Ultralytics. (2024). *YOLO11* [Computer software]. https://github.com/ultralytics/ultralytics

Upadhyay, A., Zhang, Y., Koparan, C., Rai, N., Howatt, K., Bajwa, S., & Sun, X. (2024). Advances in ground robotic technologies for site-specific weed management in precision agriculture: A review. *Computers and Electronics in Agriculture, 225*, 109363. https://doi.org/10.1016/j.compag.2024.109363

Velumani, K., Lopez-Lozano, R., Madec, S., Guo, W., Gillet, J., Comar, A., & Baret, F. (2021). Estimates of maize plant density from UAV RGB images using Faster-RCNN detection model:

Impact of the spatial resolution. *Plant Phenomics, 2021*, 9824843. https://doi.org/10.34133/2021/9824843

Wallinga, J., Groeneveld, R. M. W., & Lotz, L. A. P. (1998). Measures that describe weed spatial patterns at different levels of resolution and their applications for patch spraying of weeds. *Weed Research, 38*(5), 351-359. https://doi.org/10.1046/j.1365-3180.1998.00106.x

Wang, A., Peng, T., Cao, H., Xu, Y., Wei, X., & Cui, B. (2022). TIA-YOLOv5: An improved YOLOv5 network for real-time detection of crop and weed in the field. *Frontiers in Plant Science, 13*, 1091655. https://doi.org/10.3389/fpls.2022.1091655

Woebbecke, D. M., Meyer, G. E., Von Bargen, K., & Mortensen, D. A. (1995). Color indices for weed identification under various soil, residue, and lighting conditions. Transactions *of the ASAE, 38*(1), 259-269. https://doi.org/10.13031/2013.27838

Zingsheim, M. L., & Döring, T. F. (2024). What weeding robots need to know about ecology. *Agriculture, Ecosystems & Environment, 364*, 108861. https://doi.org/10.1016/j.agee.2023.108861