NatureNet DataHub: The Data It Takes to Field an Operational Buckthorn Detector

> We built a working drone-to-map system for invasive buckthorn and measured how close it is to operational. The detector works on the flights it trained on and collapses on a new one. Adding more labeled images of the same flights does not fix it, and the program has labeled only two flights so far. Chasing data broadly is expensive and easy to waste. The better move is a short, specific sequence of flights, cheapest first, measured after each, with a built-in test that says at every step whether the detector is ready, or whether the problem runs deeper and the work should stop.

The problem

Common buckthorn (*Rhamnus cathartica*) is an aggressive woody invasive across the Upper Midwest. It crowds out native understory plants and forms dense thickets, and managers need a way to find it that scales beyond walking the ground. Low-altitude drones make individual plants visible at sub-decimeter detail at modest cost (Singh et al. 2024), so the raw imagery is not the bottleneck.

What a manager actually needs is a detector that works on the *next* flight, not just the one it learned from. A model trained on one drone survey, with its particular light, season, sensor, and calibration, can do well there and still fail on a new survey. Because a missed infestation costs a program more than a false alarm a reviewer dismisses in seconds, the real target is reliability on flights the model has not seen.

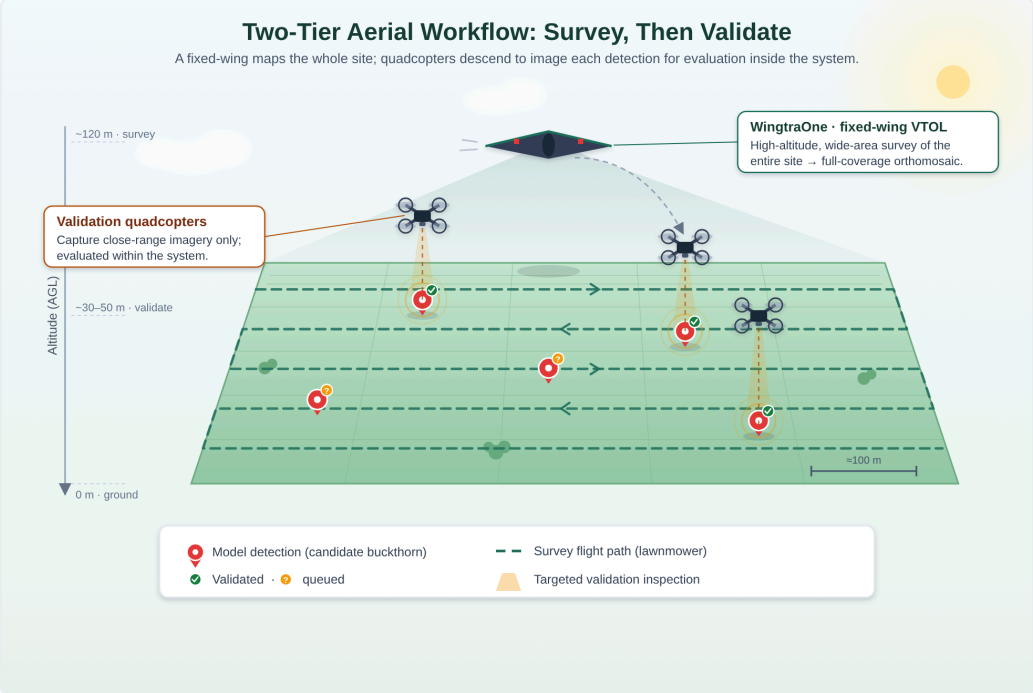

Figure 1. How the imagery is collected. A fixed-wing drone maps the whole site at altitude to build one full-coverage image, then quadcopters drop down to photograph individual candidate plants up close for review. The aircraft only capture imagery; every detection decision is made inside the system. The new flights requested below are surveys of this kind.

What we measured

We tested the detector the way it would actually be used: trained on one survey mission and evaluated on a *different* mission it had never seen (Figure 2). On a flight it trained on, the detector is workable; on a held-out flight, its accuracy drops by roughly an order of magnitude, and it flags almost none of the plants that are there. The collapse holds in both directions and across repeated training runs, so it is not a fluke or a tuning accident.

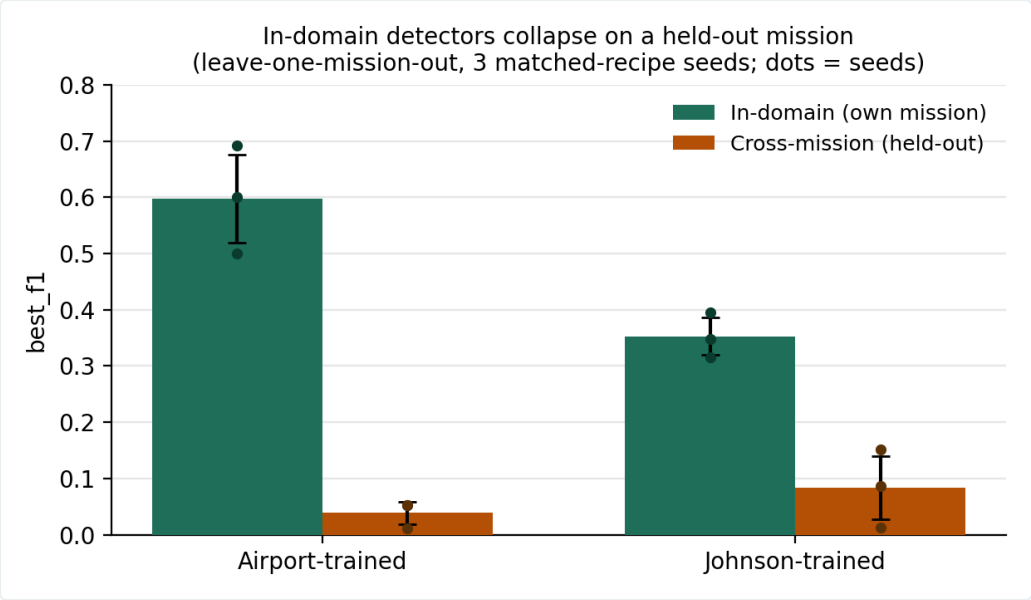

Figure 2. The same detector, scored on a flight it trained on (left bars) and on a held-out flight (right bars), for two missions and three training runs each. Workable in-flight performance collapses to near zero across flights.

One detail in the same chart points straight at the budget question. The two flights were not photographed equally closely: one had far more labeled images than the other, roughly 80 against 30, and it gave the stronger result on its own flight (the taller left bar). Yet that better-covered flight was the one whose detector did worst on a flight it had never seen (the lowest right bar). More pictures of a single flight made the detector look better at home without making it any readier for the next one.

Counted plant by plant, the same collapse is plain (Figure 3): on its own flight the detector finds 14 of 22 buckthorn; on a flight it has never seen, it finds 1 of 37.

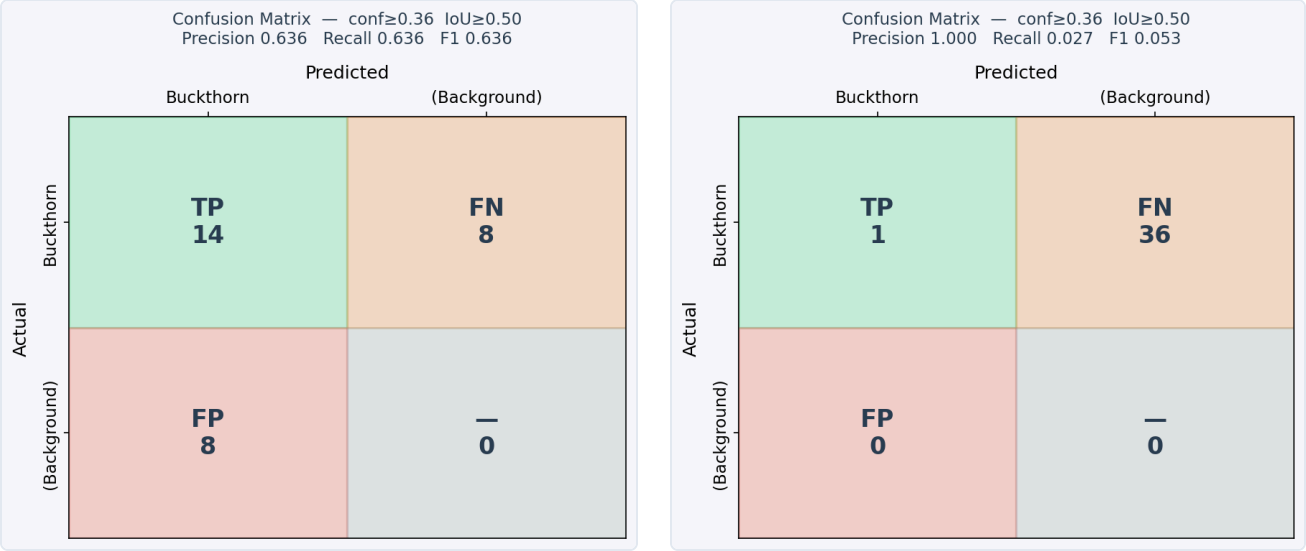

Figure 3. The same detector, counted plant by plant. Left, on a flight it trained on: 14 of 22 plants found (green), 8 missed (orange), 8 false alarms (red). Right, on a flight it has never seen: 1 of 37 found. Green is correct detections, orange is missed plants, red is false alarms.

We then asked the budget question directly, by testing what kind of "more data" would help:

- **More labels of the same flights did not fix it.** We roughly doubled the labeled images for one mission. Accuracy on that flight improved, but accuracy on a held-out flight did not move.
- **Fancier imagery did not fix it.** Adding near-infrared and red-edge bands (light beyond ordinary color, Figure 4) helped on one mission's own flight but gave no cross-flight benefit.
- **Reprocessing the imagery did not fix it.** We cut each image into smaller pieces so the plants look bigger to the detector, a standard trick for small targets. Accuracy did not change.
- **What was missing was data from other flights.** We have labeled buckthorn on only two survey missions, so the detector has never seen the range of conditions real deployment spans.

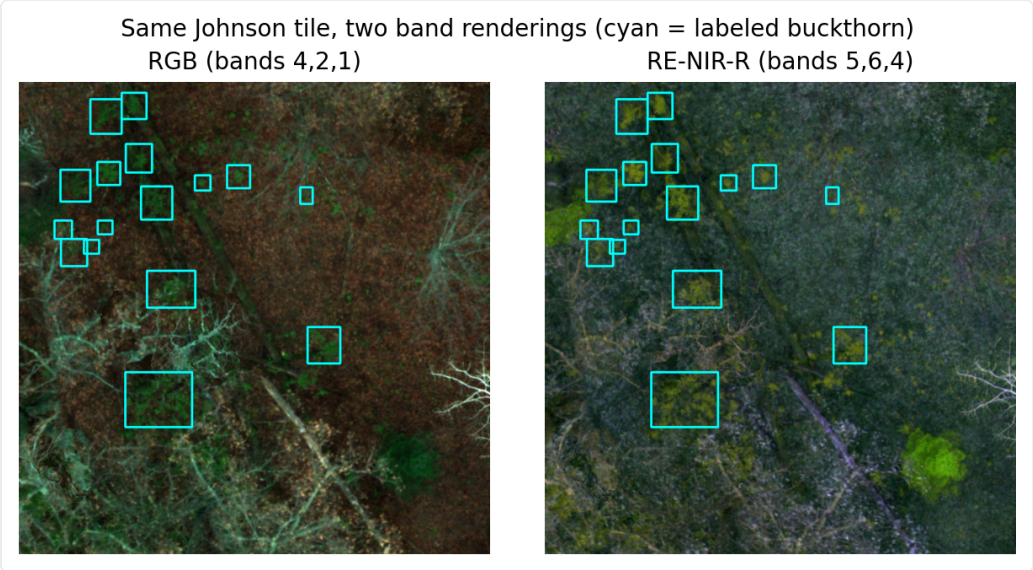

Figure 4. The same patch of ground in ordinary color (left) and in a near-infrared and red-edge false-color rendering (right); cyan boxes mark known buckthorn. The extra bands change what is visible, but on a held-out flight they gave no detection benefit.

For budgeting, there are two kinds of “more data,” and they are not interchangeable. More labels *per flight* raise accuracy on that flight but not across flights; data from *more flights* is the one lever left to test, and the surveys below are designed to show whether it delivers. And a detector evaluated only on its own flights looks nearly ready when it is not, so spending on more labels can show progress on paper while operational readiness stands still.

The full evaluation protocol, all training seeds and confusion matrices, the controls that rule out simpler explanations, and an independent integrity audit are set out in our companion methods report (Coleman, in preparation; see references).

What we are doing about it

We deployed the best available detector as a *starter model*, tuned to favor catching plants over avoiding false alarms, and built the system so that using it generates exactly the data we need: more flights, ground-truthed.

The system closes a loop: the starter model flags candidate plants, a person reviews and confirms them on a map, confirmed locations become an optimized return-flight plan that loads onto a drone controller, crews re-fly and ground-truth the sites with a GPS-aware mobile tool, and every confirmation becomes a labeled example for the next training round (Figure 5). Each pass adds what the detector lacks: data from new flights.

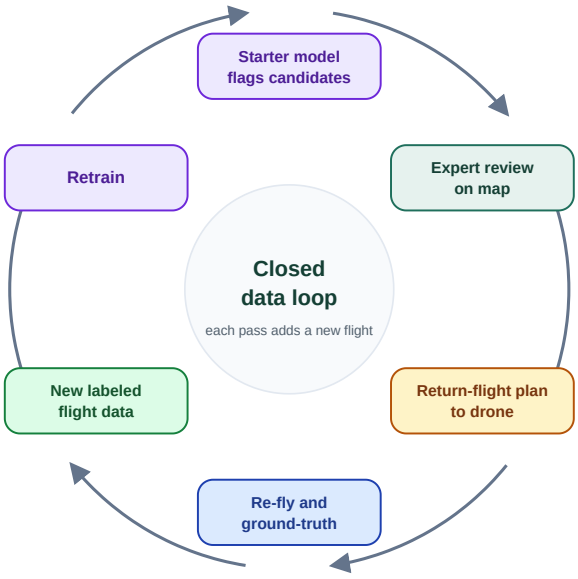

Figure 5. The closed data loop. The starter model is not the end product; it is the engine that collects data from new flights, which is what an operational detector needs.

The path forward

Because more labels of the same flights do not move cross-flight accuracy, the lever still to test is flights, and the value of a survey is the new conditions it adds. The next surveys are chosen to spread labeled ground truth across the conditions that vary between flights, season, site, sensor, and plant scale, in priority order:

1. **Re-fly a site we have already labeled.** Flying ground with known plant locations gives the same plants under new conditions: our first true cross-flight test and a matched training pair. Because the plants are already labeled, it adds

a new flight with no new annotation work, which makes it the cheapest way to buy a first cross-flight measurement, and the reason it comes first.

2. **A late-season flight.** One survey timed to late October or early November, when native trees have dropped their leaves but buckthorn stays green, adds a new flight and targets buckthorn's most distinguishable window; flight timing can matter as much as sensor quality for invasion monitoring (Müllerová et al. 2017).
3. **Enough additional surveys to measure the curve.** With three or more independent flights we can measure how cross-flight accuracy improves as flights are added, and determine the minimum data needed to reach an operational detector, or learn early if the problem runs deeper.

This is a sequenced, capped plan, not an open-ended data campaign. Each new flight is scored honestly, by training without it and testing on it, so every survey returns a real readiness number and a clear go/no-go, and the program stops or pivots the moment the gain flattens rather than spending on more data blindly. What it takes is flights and the crew time to ground-truth them; what it returns is measurable progress toward an operational tool, with a discipline that reports whether the effort is working.

Selected references

Al-Emadi, S. A., Yang, Y., & Ofli, F. (2025). Benchmarking Object Detectors under Real-World Distribution Shifts in Satellite Imagery (RWDS). *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*. arXiv:2503.19202.

Müllerová, J., Brůna, J., Bartaloš, T., Dvořák, P., Vítková, M., & Pyšek, P. (2017). Timing Is Important: Unmanned Aircraft vs. Satellite Imagery in Plant Invasion Monitoring. *Frontiers in Plant Science*, 8, 887. https://doi.org/10.3389/fpls.2017.00887

Coleman, J. (2026). *Developing Best Practices for Mission-Disjoint Evaluation of Small-Data UAV Invasive-Plant Detection.* NatureNet DataHub, companion methods report, in preparation.

Singh, K. K., Surasinghe, T. D., & Frazier, A. E. (2024). Systematic review and best practices for drone remote sensing of invasive plants. *Methods in Ecology and Evolution*, 15(6), 998-1015. https://doi.org/10.1111/2041-210X.14330

Quantitative claims trace to primary run artifacts under …`/ml-training/output/` through `WHITE_PAPER_FACTS.md`: the cross-flight collapse (in-domain best_f1 ~0.60 falling to ~0.04 on a held-out mission, three matched-recipe seeds; "Matched-recipe MULTI-SEED LOMO" row, with Figure 2 from `output/lomo-ms/`); more in-domain labels lifting in-domain but not cross-flight accuracy ("Johnson model rebuilt" row); multispectral giving no cross-flight benefit ("Multispectral (RE-NIR-R) cross-flight LOMO" row); and only two labeled missions ("OOD evals are same-mission" findings). Figures 1 and 3 are schematic.