Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Small Effects, Big Noise: A Methodological Audit of Module Evaluation in Small-Scale Crack Instance Segmentation

**Xin Cao**

2012212079@qq.com

Chongqing Jiaotong University

**Article**

**Keywords:** crack instance segmentation, YOLO-seg, statistical reliability, seed variance, minimum detectable effect, negative results

**Posted Date:** November 16th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11379838/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Small Effects, Big Noise: A Methodological Audit of Module Evaluation in Small-Scale Crack Instance Segmentation

**Authors:** Xin Cao*

**Affiliations:** Chongqing Jiaotong University, Chongqing, China

*** Corresponding author** — email: 2012212079@qq.com

## Abstract

Crack instance segmentation papers built on YOLO-seg routinely claim gains by contrasting a baseline against the same model plus a proposed module, yet their statistical reliability is rarely examined. We audit a representative module, a skeleton-aware label assigner (STAL), across three epoch budgets, three seeds, three calibers, and eight metrics, and replicate on a second dataset and a second, structurally different module. Sweeping its tolerance proves the module is not inert; yet no gain is usable. The paired effect stays below 0.9 pt and flips sign with budget and caliber; the cross-seed correlation flips sign on identical runs; and the minimum detectable effect exceeds the observed one by 2.7–13.3×. All seven domain metrics are undecidable, one quantised to 11.1 pt by its 9 crossing pairs, and with three seeds exact Wilcoxon cannot return $p < 0.25$ at any effect size. Both replications reproduce the resolution ceiling, not the effect. From these results we derive an eight-rule minimum discernible protocol: effect-size floor, fixed caliber, verified pairing, MDE beside every difference, per-seed dispersion, fixed budgets, computed seed budget; every rule checkable from a single per-seed table.

**Keywords:** crack instance segmentation; YOLO-seg; statistical reliability; seed variance; minimum detectable effect; negative results

## 1. Introduction

### 1.1 Motivation

Surface cracks are among the earliest symptoms of pavement degradation, and their timely detection is a prerequisite for condition rating and preventive maintenance of transportation infrastructure [2]. Manual survey is slow, hazardous, and inconsistent across inspectors; automated inspection from vehicle-mounted cameras is therefore the field's de facto direction. Instance-segmentation models localise each crack as a pixel-level mask rather than a coarse box, preserving topology and width, which feed engineering metrics; the YOLO-seg family [6] is a popular backbone for its speed–accuracy trade-off on edge hardware.

### 1.2 The "module improvement" reporting culture

The dominant evaluation ritual is a paired comparison: a baseline detector versus the same detector augmented with a *proposed module*, claimed to raise mAP or a domain metric by a fraction of a point — attention gates, skeleton-aware label assigners, directional decoders, high-resolution prototypes, each typically reported to "improve" the baseline.

What is striking is how rarely such a claim carries the information needed to judge it: variance across seeds, the metric caliber (best-epoch vs. a windowed mean), the stopping rule, and a significance statement. A gain reported as, say, +0.25 mAP points on three seeds could be a real effect or seed-level noise pointing the same way. "The curve went up" has become evidence, without a statistical floor beneath it.

### 1.3 Why such claims are fragile

Small-scale segmentation experiments are fragile in five independent ways. **Seed variance dominates the signal**: with three seeds, the paired difference is estimated from three numbers and its

standard deviation from two degrees of freedom. **Metric caliber is a free parameter**: reporting the best epoch is implicit selection that we measure to bias comparisons by up to 9.24 mAP points. **Stopping rules are not controlled**: early stopping can terminate the two arms at different epoch counts. **Stochastic weight averaging does not behave as assumed**: meant to reduce variance, in our paired setup it amplifies the paired difference 2.04×. **The paired design assumes correlated noise**: if the cross-seed correlation ρ is near zero or negative, pairing does not cancel variance — it can inflate it. Any one can turn a null result into an apparent "improvement", or conceal a real one.

### 1.4 Our study: an audit, not a method

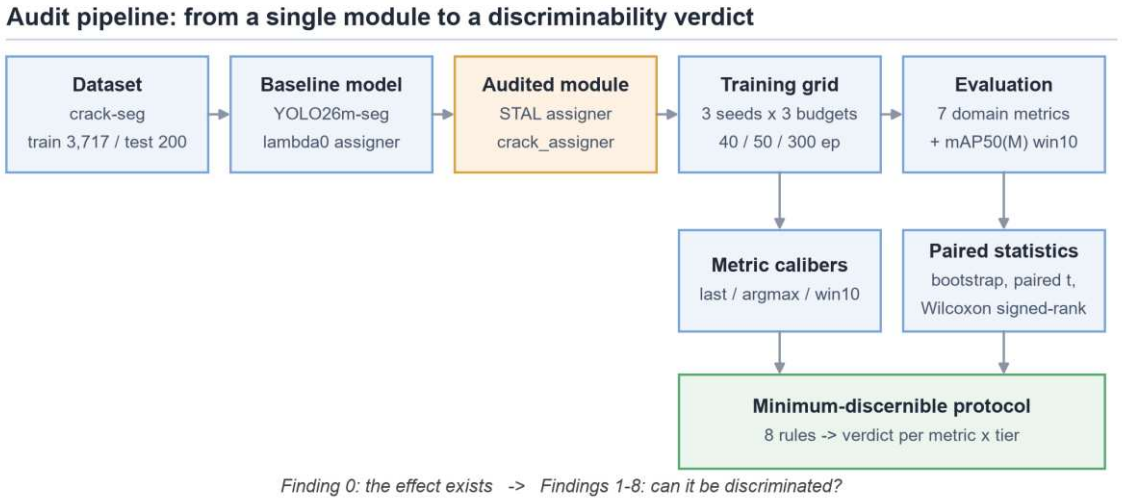
Rather than proposing another module, we make module evaluation itself the object of study. We audit one representative module — a skeleton-aware label assigner (STAL) — under a protocol specifying more than the reporting norm requires: three epoch budgets, three random seeds (42/7/2026), three metric calibers, SWA on/off, and seven domain metrics on a labelled test set of 112 images. To check that the verdict is not an artefact of this module's identity, we repeat the comparison with a second, structurally different module (soft-clDice, a loss-side topology term) and on a second dataset. We first establish that the module *does* produce an effect, then ask the harder question: **can that effect be reliably discriminated from seed noise under the small-sample regime the field actually uses?**

### 1.5 Contributions

Our contributions are: (C1) nine concrete findings (⓪–⑧) about how small-scale module comparisons behave under the field's own conventions; (C2) a minimum discernible protocol (§5) whose eight rules are each checkable in review from a single per-seed table; (C3) a second-dataset replication (CCU-Crack) showing that what travels is the *resolution ceiling*, not an effect; (C4) a clean separation of two claims routinely conflated — Finding ⓪ establishes a real effect, so everything that follows studies *discriminability*, not whether the module is inert; and (C5) a module-level control on a structurally different module (soft-clDice, loss-side) reproducing the undecidable verdict, so the audit's outcome is not an artefact of module identity.

### 1.6 Organisation

§2 reviews related work; §3 states the setup and decision rule; §4 presents the nine findings; §5 turns them into the protocol; §6 discusses implications and limitations; §7 concludes.



## 2. Related Work

### 2.1 Crack instance segmentation and its evaluation metrics

Pixel-level crack segmentation has attracted sustained attention because cracks are thin, topologically connected, and severely class-imbalanced structures whose width and continuity carry direct engineering meaning [2]. Early pipelines used fully convolutional or encoder–decoder networks; more recently, real-time one-stage YOLO-seg detectors have been adopted for their speed–accuracy trade-off, and crack-specific variants report gains over a YOLO baseline.

Two properties make evaluation unusually fragile. **First, extreme class imbalance** (fewer than 1.3 instances per image, well under 1% crack pixels) means region-overlap metrics are dominated by the easy background majority. **Second, the field has moved to topology- and instance-aware metrics** — clDice for skeleton/connectivity preservation [9] — and to benchmark families that disaggregate false responses, localisation, and instance-level merge/split behaviour [1]. These are more informative, but several depend on small instance counts and thus have limited resolution — a property we quantify in Finding ④.

### 2.2 Label assignment as a representative proposed module

Our audited module, STAL, belongs to the family of *label assignment* methods that decide which ground-truth object a prediction should learn from. Modern assigners range from static IoU thresholds [7] to adaptive, cost-based, or one-to-one strategies [8]; topology- or skeleton-guided variants are a natural specialisation for thin structures.

We chose label assignment because it is *representative*, not unusual: like attention gates or directional decoders, it is a drop-in module whose value is shown by a paired "baseline vs. baseline+module" comparison reporting a fraction-of-a-point gain — exactly the reporting pattern we scrutinise.

### 2.3 Statistical reliability of small-scale deep-learning reports

**Single-run reporting is the norm.** Picard [3] scanned up to $10^4$ seeds on CIFAR-10 and showed that even when mean variance is modest, "black-swan" seeds producing community-significant differences are easy to find, and the spread does not shrink after convergence; Bouthillier et al. [4] quantified that a meaningful fraction of reported accuracy differences vanishes once seed variance is accounted for.

**Within-algorithm differences can be "significant."** A 2024 study ran one segmentation algorithm (nnU-Net) under 50 seeds and found the best seed statistically significantly outperformed 0–76% of the others on hold-out validation [5] — direct backing for Findings ⑥/⑦: a significant p-value between two arms can arise from seed noise alone, even when the underlying algorithm is identical.

**Caliber and stopping rules are unexamined degrees of freedom.** Reporting the best epoch is implicit selection that inflates comparisons: we measure up to 9.24 mAP points of bias from this choice alone. Stochastic weight averaging is widely assumed to reduce variance by averaging near optima [10], yet in our paired setup it amplifies the paired-difference spread by 2.04× (Finding ②). Early stopping terminates arms at unequal lengths, so the same windowed metric compares different training stages (Finding ⑤).

### 2.4 Negative results and methodological audits

The community has begun valuing rigorous negative and methodological results, with dedicated tracks and audit papers that stress-test evaluation practice rather than propose methods [3, 4, 5]. Our work belongs to this lineage: we claim no new module, but a reusable protocol for deciding when a small-sample module claim can and cannot be believed.

## 3. Experimental Setup

### 3.1 Dataset and evaluation metrics

All experiments use the **crack-seg** benchmark: 3,717 training, 200 validation, and 200 test images, of which **112 carry instance-level annotations** and are used for all domain-metric evaluation with this study's suite (FR / SLR / SLP / MEI / clDice / HI / IMR; implemented in `crack_topology_metrics.py`, localisation tolerance tol = 2 px). HI and MEI follow the length-based caliber of Perri et al. (Appl. Sci. 2026, 16(17):8608); FR, SLR, SLP and IMR are defined here and are **not** claimed to be a published benchmark. As is typical for crack imagery, foreground pixels occupy well under 1% of the image area, so region-overlap metrics are insensitive to the errors that matter operationally.

*(Dataset statistics: Supplementary Table 1.)*

We therefore report the primary detector metric **mask mAP50(M)** together with **seven domain metrics**: four where higher is better — SLR and SLP (skeleton-length and skeleton-pixel recall), **clDice** [9] (topological connectivity), and mAP50(M) — and three where lower is better — **FR** (false-response rate), **MEI** (missed-error instances), and **HI** (hallucinated length). We also report **IMR**, the *instance merge rate*: among the **9 ground-truth crossing pairs** in the test set (neighbouring cracks a detector may erroneously merge into one instance), IMR is the fraction the model merges. IMR is included precisely because it is the metric a reviewer would propose when a module claims "better separation of adjacent cracks"; §4 shows it has the least resolution.

To test whether the conclusions are specific to one benchmark, we repeat the ground-truth statistics, and later the full comparison, on a second independently captured dataset, **CCU-Crack** (Supplementary Table 2).

### 3.2 Baseline and audited module

The baseline (**lambda0**) is **YOLO26m-seg** trained from scratch (epoch-1 mAP50(M): 0.129 vs. 0.111 for the module arm; both arms share the same pre-trained backbone configuration). The audited module, **STAL**, is a skeleton-aware label assigner (`CrackSkeletonAssigner`) whose single behavioural switch is `crack_assigner` (true/false) with axial tolerance σ (default 0.5); when the assigner is disabled, σ has no effect. As a **module-level control** (§4.11) we also train a third arm with **soft-clDice** [9] — a topology-preserving *loss* term acting on the gradient rather than on sample-to-ground-truth allocation — at weight λ = 0.05; its baseline arm is the same λ0 configuration, so the control reuses the crack-seg 40-ep baseline runs at no extra cost.

**Configuration audit (Supplementary Table 3).** Because every paired comparison's credibility rests on the two arms differing *only* in the audited module, we audited every differing field in the training configurations across all 18 paired runs. Exactly one field distinguishes the arms: `crack_assigner`. All other differences are benign and verified (Supplementary Table 3); the dataloader `workers` item we controlled empirically: a dedicated paired run at 40 ep / seed 42 with workers changed from 0 to 8 shifted win10 mAP50(M) by only **−1.12 pt** (argmax −0.91 pt) — an order of magnitude below the effects under audit and opposite in sign to our initial (incorrect) suspicion; workers is therefore a non-confound.

### 3.3 Training protocol

Each arm is trained under **three epoch budgets** — 40, 50, and 300 epochs — with **three random seeds** (42, 7, 2026), batch size 16, image size 640, and the optimizer's automatic schedule (resolved to AdamW, lr = 0.002 in the audited runs). The 300-ep budget is an *upper* budget with early stopping (patience = 30); in practice **all six 300-ep runs stopped early, between 54 and 145 epochs**, and the two arms of the same seed could differ by up to **35 epochs** (Finding ⑤). The 40- and 50-ep budgets ran to completion and are treated as fixed-length.

### 3.4 Evaluation calibers

From each run's validation curve we compute three **calibers**: the last epoch (**last**), the best epoch (**argmax**), and the mean over the final 10 epochs (**win10**), adopted as the **primary caliber**. The argmax caliber is reported for comparison only: it inflates the apparent baseline-vs-module gap by up to **9.24 mAP points** through implicit selection of a favourable epoch, and its paired cross-seed correlation has the *opposite sign* to win10 (Finding ①). For the domain metrics (Supplementary Tables 7 and 8) each model's **final-epoch checkpoint (`last.pt`)** is used; the SWA snapshots saved alongside the 50-ep runs were **not** used (the `swa50_` prefix on the 50-ep λ0 run directories reflects this — see the Supplementary Table 6 footnote).

### 3.5 Decision rule

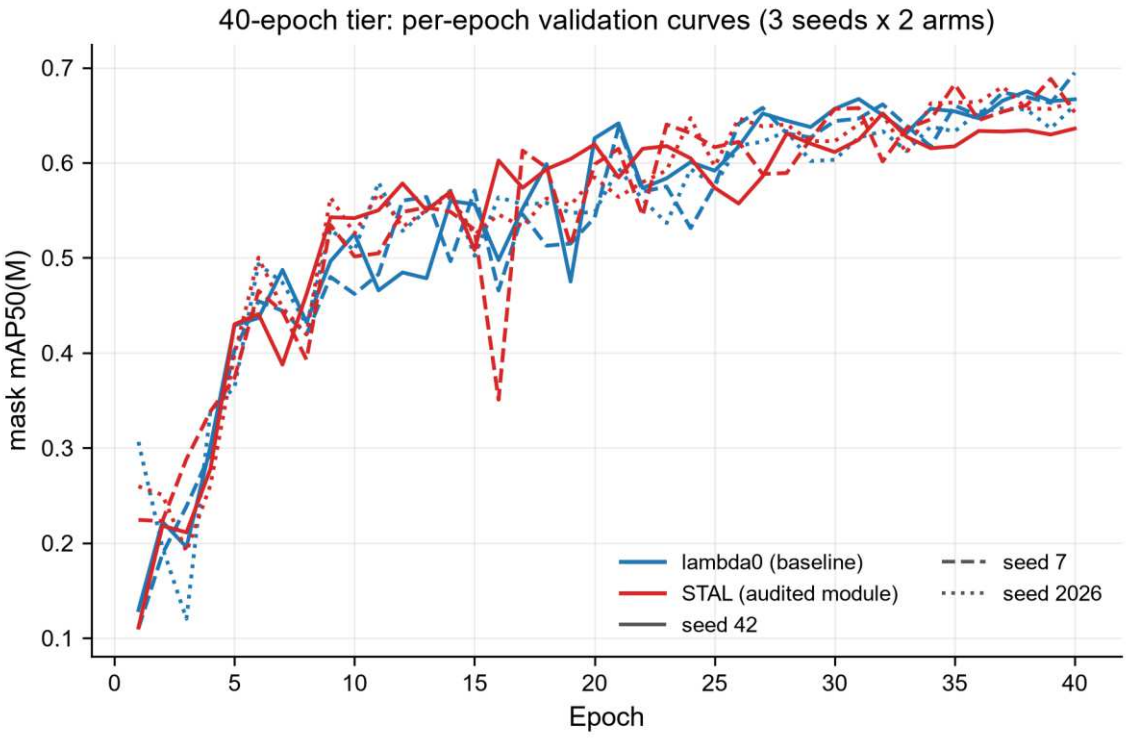
For every metric we form the seed-paired difference Δ = STAL − lambda0 and declare the effect **discernible** only if *all* of the following hold: (i) $|\Delta| \geq 2$ pt (an effect-size floor, chosen because the median minimum-detectable effect at n = 3 is ≈3.5 pt); (ii) the two-sided paired t-test reaches $p < 0.05$; and (iii) the bootstrap 95% CI of Δ excludes zero. The Wilcoxon signed-rank test is reported for completeness but is structurally incapable of significance at n = 3 (minimum attainable two-sided p = 0.25, Finding ⑥). When the criteria are not met we report **UNDECIDABLE** rather than "no difference": the protocol separates "the module has no effect" (rejected a priori by the σ-sweep, Finding ⓪) from "the effect cannot be discriminated at this sample size".

### 3.6 Implementation and reproducibility

All runs execute on a single consumer GPU (Windows, PyTorch/YOLO26); epoch times range from ≈83 to ≈192 s depending on budget and assigner state. Every number in the paper is recomputable from the released per-epoch validation curves alone, and the seeds, budgets, and decision rule were fixed before the domain-metric evaluation was performed; the release itself is described in the Data Availability section.

## 4. Findings

All differences are Δ = STAL − λ0 in percentage points (pt), paired by seed, under win10 unless stated.

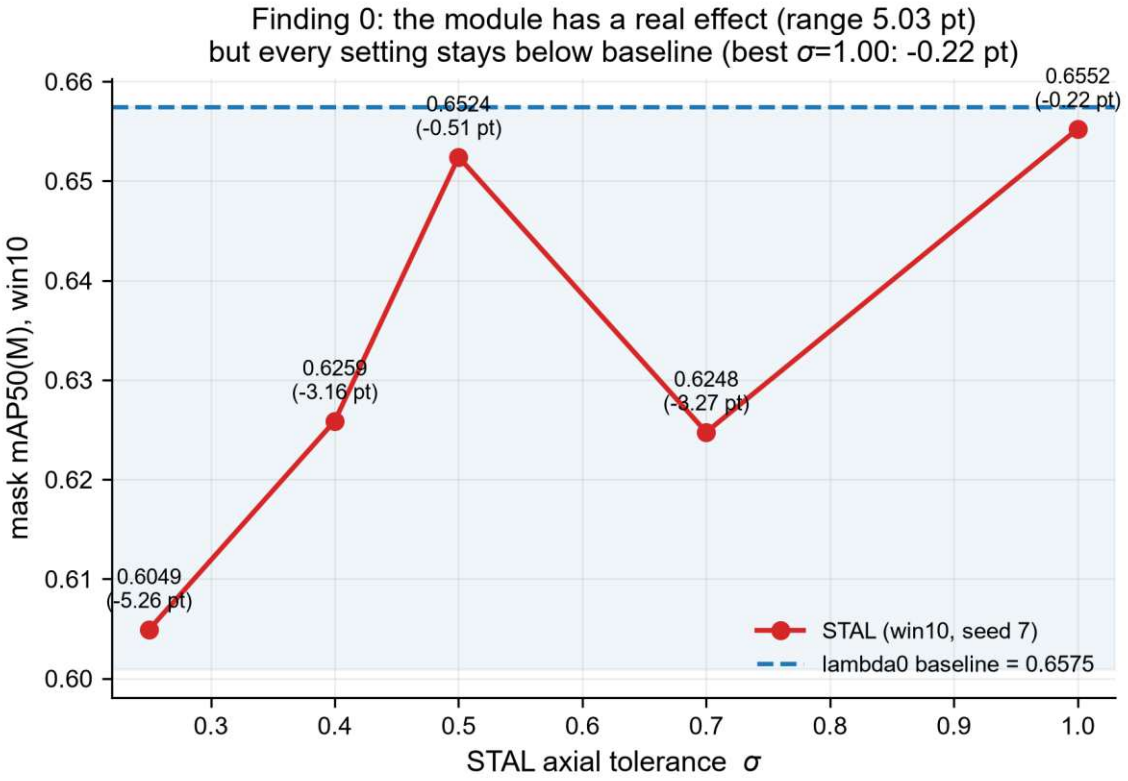


| Caliber | λ0 (s42 / s7 / s2026) | STAL (s42 / s7 / s2026) | Δ per seed (pt) | mean Δ (pt) | sd(Δ) | t (df=2) | p | ρ | MDE |
|---|---|---|---|---|---|---|---|---|---|
| **win10 (primary)** | 0.65785 / 0.65747 / 0.64063 | 0.63022 / 0.65242 / 0.65494 | −2.764 / −0.505 / +1.430 | **−0.613** | **2.099** | −0.506 | 0.663 | **−0.594** | 5.21 |
| last | 0.66680 / 0.69483 / 0.66074 | 0.63602 / 0.65309 / 0.66449 | −3.078 / −4.174 / +0.375 | −2.292 | 2.374 | −1.672 | 0.236 | −0.053 | 5.90 |
| argmax | 0.67519 / 0.69483 / 0.66074 | 0.65196 / 0.68828 / 0.68003 | −2.323 / −0.655 / +1.929 | −0.350 | 2.142 | −0.283 | 0.804 | +0.301 | 5.32 |

The base tier already shows the pattern the rest of §4 dissects: the mean is small, the per-seed scatter is large (sd(Δ) = 2.10 pt), individual seeds cross the 2 pt floor in both directions (−2.76 pt and +1.43 pt), and the detectable effect (MDE = 5.21 pt) is 8.5× the observed one.

### 4.0 Finding ⓪ — Real effect, not a usable gain (Supplementary Table 4, Fig. 3)

Sweeping σ over {0.25, 0.40, 0.50, 0.70, 1.00} at 40 ep / seed 7 moves win10 mAP50(M) across a **5.03 pt range** — unambiguous evidence the module changes behaviour — yet all five settings fall below the baseline (0.657472). The best still trails it, and the response is non-monotone (σ = 0.70 is worse than 0.50). **Implication:** "the module does nothing, hence nothing is detected" is rejected.

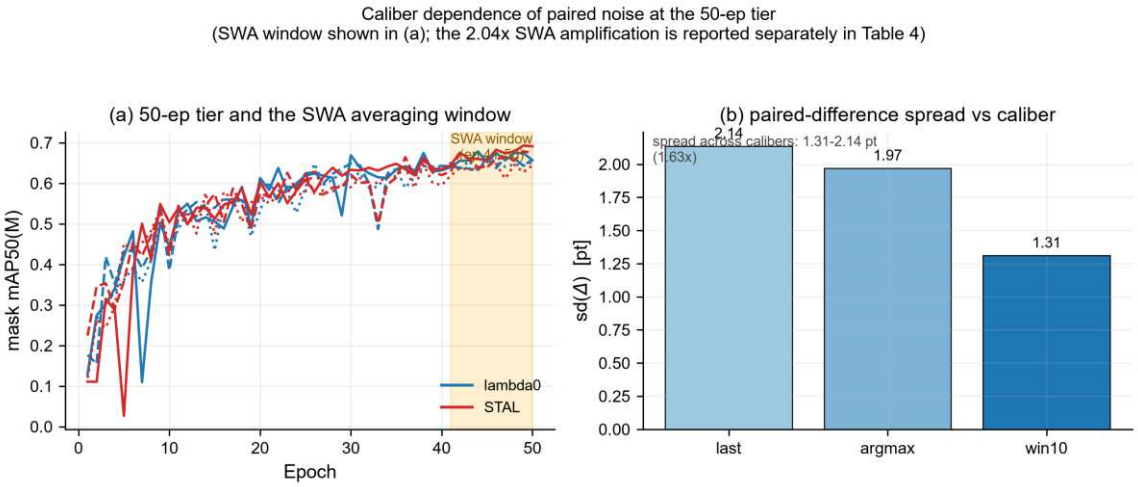



### 4.1 Finding ① — ρ is caliber-dependent and flips sign (Supplementary Table 5, Supplementary Fig. 1)

The same six runs yield ρ = +0.892 under win10 and −0.999 under argmax. A negative ρ means paired differencing does not cancel seed variance — it amplifies it.

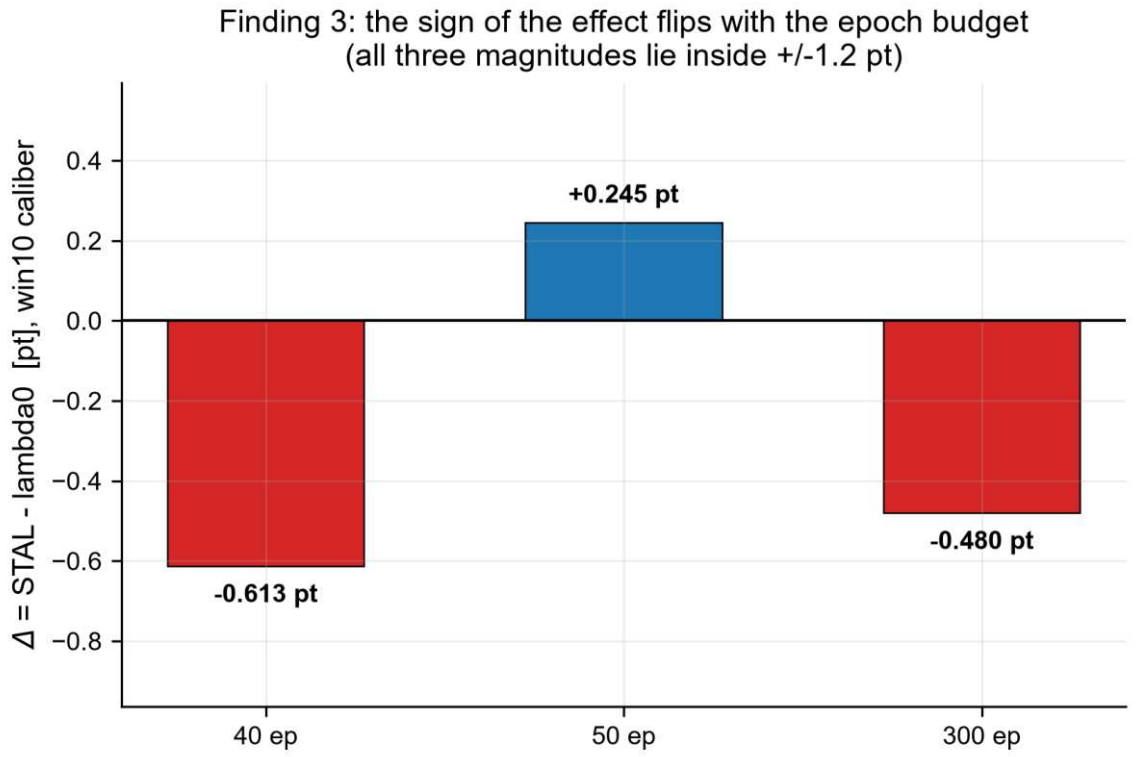
### 4.2 Finding ② — Caliber dominates the spread; SWA amplifies rather than reduces (Supplementary Table 6, Fig. 4)

At 50 ep, sd(Δ) is **2.14 / 1.97 / 1.31 pt** for last / argmax / win10 — a 1.63× spread from how the scalar is read off the curve. Evaluating all six SWA-averaged checkpoints offline gives per-seed Δ of +1.401 / +0.049 / −3.761 pt: **sd(Δ) = 2.677 pt, a 2.04× amplification over win10** — the opposite of the intended effect — and it reverses the mean effect's sign (−0.770 pt vs. +0.245 pt under win10).




### 4.3 Finding ③ — The sign flips with the budget (Fig. 5)

Δ = **−0.613 (40 ep), +0.245 (50 ep), −0.480 pt (300 ep)**; p = 0.66 / 0.78 / 0.52. Across all budgets no *mean* Δ exceeds 0.62 pt in absolute value. Individual seeds do exceed the floor within a budget (e.g. Δ = −2.76 pt at one seed of the 40-ep tier, Table 1); dispersion and central tendency tell different stories, which is why §5.6 requires both. The sign of a sub-point effect is not a stable property of a module but of the budget at which the curve happened to be read.



### 4.4 Finding ④ — Seven domain metrics undecidable; IMR has almost no resolution (Supplementary Tables 7 and 8; Supplementary Fig. 2)

Across both tiers no domain metric satisfies all three decision criteria; the largest |mean Δ| is 0.92 pt at 50 ep (HI) and 1.14 pt at 300 ep (SLP).

IMR is the limiting case. With only **9 crossing pairs**, the crossing-merged rate is quantised to multiples of 1/9 (11.1 pt). At 50 ep all six models return exactly 0.1111 (sd = 0 — a degenerate "perfect" agreement); at 300 ep IMR takes only two values, {0, 0.1111}, flipping across seeds for **sd(Δ) = 11.11 pt and MDE = 27.6 pt**: a resolution quantum of 11 pt cannot register a sub-point improvement.

**A live example of small-sample inconsistency.** At 300 ep, SLP's bootstrap CI is [−2.00, −0.29] pt — it *excludes* zero — while the paired t-test gives p = 0.147 and |Δ| = 1.14 pt < 2 pt. Two standard tools disagree on the same three numbers; we follow the pre-registered rule and report UNDECIDABLE — the disagreement itself is the phenomenon this paper is about.

### 4.5 Finding ⑤ — Early stopping leaves the arms at unequal lengths (Supplementary Table 9, Supplementary Fig. 3)

All six 300-ep runs early-stopped at **54 / 145 / 118 epochs (λ0)** and **89 / 116 / 117 (STAL)** for seeds 42 / 7 / 2026 — a same-seed gap of up to **35 epochs**, so one arm's win10 window is centered at a different training stage than the other's.

### 4.6 Finding ⑥ — Nonparametric significance is structurally impossible at n = 3

The exact Wilcoxon signed-rank test with n = 3 attains, in the most favourable case, two-sided **p = 2 × (1/8) = 0.25** — a combinatorial bound depending on no distributional assumption and no effect size: with three seeds, Wilcoxon can never return $p < 0.05$, however large or consistent the effect. Consistently, the smallest $p_W$ observed anywhere is 0.50. "We ran three seeds and it was not significant" is not evidence about the module but a tautology about the design.

### 4.7 Finding ⑦ — The detectable effect is 4.8–13.3× the observed effect (Supplementary Fig. 4)

$\mathrm{MDE} = t_{0.975,2} \cdot \mathrm{sd}(\Delta)/\sqrt{3} = 4.303 \cdot \mathrm{sd}(\Delta)/1.732 \approx 2.48 \times \mathrm{sd}(\Delta)$ at n = 3.

The ratio of detectable to observed effect ranges from **4.8× (SLP) to 13.3× (FR, mAP50(M))** across metrics, and the seed budget required to close it is **17–113** (35–231 at 80% power). *(Per-metric values: Supplementary Table 12.)*

### 4.8 Finding ⑧ — Pairing fails at short budgets but recovers at 300 ep

At 50 ep, ρ is negative for five of six non-degenerate metrics (SLR −0.068, SLP −0.539, MEI −0.046, clDice −0.422, HI −0.786); only FR is positive (+0.909). At 300 ep, five of six are positive (clDice +0.994, FR +0.934, SLP +0.720, SLR +0.693, MEI +0.669), with only HI negative (−0.981).

Pairing failure is therefore a *short-budget* phenomenon. Critically, even where pairing succeeds at 300 ep, n = 3 still cannot reach significance (lowest $p_t = 0.147$) — the binding constraint is ⑥, not the pairing design.

### 4.9 Cross-tier summary

Under the primary caliber the effect never exceeds 0.62 pt, its sign changes with budget, and it is undecidable on mAP and all seven domain metrics at both 50 and 300 ep. The negative result is therefore not an artefact of any single training length, module, or benchmark: replacing the module with a structurally different one (§4.11) or the dataset with a second one (§4.10) reproduces the same pattern — an observed effect smaller than the detectable one.

### 4.10 Cross-dataset check (CCU-Crack)

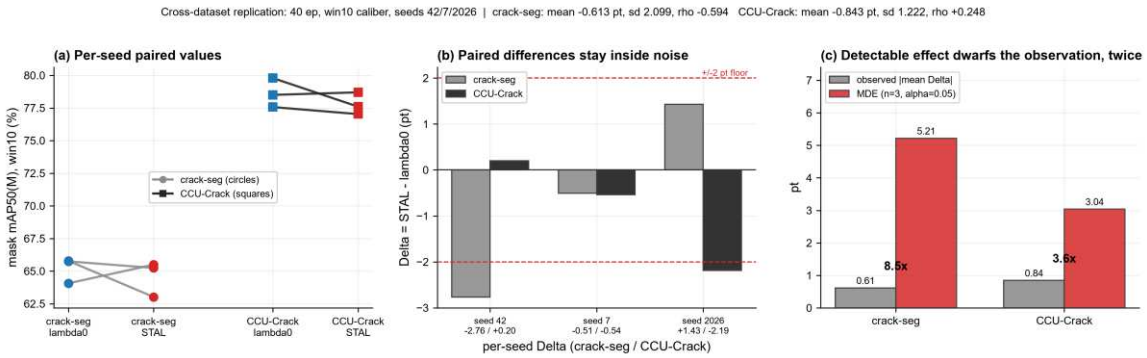
We repeat the analysis on a second, independently captured dataset, CCU-Crack (1,216 train / 268 val; Supplementary Table 2).

**Resolution collapse generalises.** CCU-Crack's validation set contains only **5 crossing pairs**, so IMR is quantised to 1/5 = **20.0 pt**, nearly twice as coarse as crack-seg's 11.1 pt: metrics depending on rare topological events are unusable for sub-point effects on *both* benchmarks. Finding ④ is a cross-dataset property, established at zero training cost.

**Replication of the paired comparison.** We ran the full 2 arms × 3 seeds × 40 ep comparison on CCU-Crack at a **fixed budget (patience = 40, no early stopping)**, so both arms of every seed have equal length, avoiding the Finding ⑤ problem of the 300-ep tier. All six runs completed 40/40 epochs; every number below was recomputed from the raw `results.csv` files.

**Pre-registered prediction (stated before the run): |Δ_win10| < 2 pt, sd(Δ) ≥ 1 pt, verdict UNDECIDABLE — all three confirmed.** Observed |mean Δ_win10| = **0.843 pt** (< 2 pt ✓); sd(Δ) = **1.222 pt** (≥ 1 pt ✓); p = 0.355, and by Finding ⑥ an n = 3 design cannot reach significance at any effect size, so the verdict is **UNDECIDABLE** ✓.

**What the replication adds.** (i) The sign of Δ is negative on CCU-Crack as on crack-seg (40 ep win10: mean Δ = −0.613 pt): the direction is reproduced, though neither is significant. (ii) The detectable effect again dwarfs the observation: MDE = **3.04 pt** vs |mean Δ| = 0.843 pt, a factor of **3.6×** — Finding ⑦ transfers, at a ratio just below the 4.8–13.3× range established on crack-seg. (iii) ρ again fails to reproduce: ρ_win10 = **+0.248** on CCU-Crack vs **−0.594** on crack-seg at the same budget and caliber, while the *same runs* give ρ = +0.946 (last) and +0.938 (argmax) — a quantity that swings from −0.59 to +0.25 across datasets and from +0.25 to +0.95 across calibers is an artefact of the measurement design, not of the module (Finding ①/⑧ on new data).




**Why a single seed exceeding the floor does not overturn the verdict.** Seed 2026 gives Δ = −2.19 pt (Supplementary Table 10), crossing the 2 pt floor — but the decision rule applies all three criteria jointly to the *paired mean*, not to a single seed; with sd(Δ) = 1.22 pt, per-seed scatter beyond the mean is expected, and reporting it is what the protocol asks.

**Honest qualification.** By our own Finding ⑥, this n = 3 replication cannot support a significance claim either: it replicates the *variance structure and the undecidability*, not an effect, and we say so rather than presenting it as confirmatory evidence.

### 4.11 Module-B control: a second, structurally different module

**Why a control module.** A reviewer can reasonably ask whether UNDECIDABLE is a property of *this* module rather than of the measurement regime. We therefore run a **module-level control** with a structurally different second module: **soft-clDice** [9], a topology-preserving *loss* term (a soft harmonic mean of Dice overlap and skeleton recall) acting on the gradient, where STAL acts on prediction-to-ground-truth allocation. If the verdict survives the swap, it is not about module identity.

**Design.** Same dataset (crack-seg) and protocol as the 40-ep tier: 2 arms × 3 seeds (42/7/2026) × **40 ep fixed budget** (patience = 40, no early stopping), same three calibers, same decision rule (§3.5). The

λ0 arm is the existing 40-ep baseline (§3.2), so the control costs three training runs, not six; the only differing field is `cldice_weight` = 0.05. All three runs completed 40/40 epochs; every number below was recomputed from the raw `results.csv` files.

**Pre-registration and stop rule (stated before any λ = 0.05 data existed).** Prediction: |mean Δ_win10| < 2 pt, sd(Δ) ≥ 0.5 pt, verdict **UNDECIDABLE**. Stop rule: if ≥ 2 of 3 seeds diverge at the chosen weight, declare the configuration untrainable and stop — the weight is *not* lowered again, so a flattering run cannot be fished out of a λ scan.

**Attempt log (reported because a negative result that hides its failed attempts is not a negative result).** At λ = 0.1 the loss diverged on **all three seeds** — the finite guard aborted training in epochs 31 / 29 / 27 after 30 / 28 / 26 completed epochs, the clDice gradient term drove the prototype head into fp16 overflow. We record this as an *instability* finding, not an ineffectiveness finding: at half the weight the module trains cleanly and is merely undecidable. A module whose reported weight crashes training three times out of three would be dismissed as "does not work"; fragility to one hyperparameter is part of what small-scale comparisons hide.

| Caliber | λ0 (s42 / s7 / s2026) | soft-clDice (s42 / s7 / s2026) | Δ per seed (pt) | mean Δ (pt) | sd(Δ) | t (df=2) | p | ρ | MDE |
|---|---|---|---|---|---|---|---|---|---|
| **win10 (primary)** | 0.65785 / 0.65747 / 0.64063 | 0.64821 / 0.64561 / 0.64224 | −0.964 / −1.186 / +0.161 | **−0.663** | **0.723** | −1.590 | 0.253 | **+0.909** | 1.80 |
| last | 0.66680 / 0.69483 / 0.66074 | 0.66581 / 0.65008 / 0.67207 | −0.099 / −4.475 / +1.133 | −1.147 | 2.947 | −0.674 | 0.570 | −0.994 | 7.32 |
| argmax | 0.67519 / 0.69483 / 0.66074 | 0.67940 / 0.66878 / 0.67207 | +0.421 / −2.605 / +1.133 | −0.350 | 1.985 | −0.306 | 0.789 | −0.385 | 4.93 |

**All three pre-registered predictions are confirmed.** Observed |mean Δ_win10| = **0.663 pt** (< 2 pt ✓); sd(Δ) = **0.723 pt** (≥ 0.5 pt ✓); p = 0.253, and by Finding ⑥ an n = 3 design cannot reach significance at any effect size, so the verdict is **UNDECIDABLE** ✓. The secondary calibers agree: mean Δ = −1.147 pt with MDE 7.32 pt (last), −0.350 pt with MDE 4.93 pt (argmax) — undecidable on all three.

**What the control adds.**

*(i) The verdict is module-identity independent.* A loss-side and an assignment-side term both land below the detectable floor (MDE = **1.80 pt** vs. |mean Δ| = 0.663 pt, **2.7×**): what prevents resolving STAL's effect is the measurement regime — n = 3, one caliber, one budget — not the module. This converts the "one module" limitation into a two-module result.

*(ii) Pairing was legitimate here — and still failed.* The STAL arm shows ρ_win10 = −0.594, where paired differencing is statistically inadmissible; the control shows **ρ = +0.909**, where it is admissible. The comparison is undecidable either way: even with the paired design's central assumption verified, three seeds cannot resolve a sub-point effect — hence rules stated in terms of the MDE, not the available test.

*(iii) ρ is caliber-dependent again, on identical runs.* +0.909 (win10), −0.994 (last), −0.385 (argmax): the paired correlation's sign swings with the reading convention on the *same three runs* — Finding ① restated on a second module.

**Honest qualification.** This control is still n = 3, one dataset, one network family. It does not estimate soft-clDice's true effect and we make no claim that it does; it demonstrates that the audit's verdict is reproducible when the module is swapped, the only claim it needs to support.

## 5. A Minimum Discernible Protocol

**5.1 Effect-size floor.** Require $|\Delta| \geq 2$ pt before any significance claim. At n = 3 the median MDE is ≈3.5 pt.

**5.2 Fix the caliber in advance; never report argmax.** Use a windowed mean over the final k epochs (k = 10); argmax carries up to **9.24 pt** of selection bias and can invert ρ's sign.

**5.3 Prefer independent two-sample comparison; pair only when ρ > 0 is verified** per metric and caliber.

**5.4 Report interval and test together; expect disagreement at small n** (cf. SLP at 300 ep).

**5.5 Pre-register the decision rule; publish the MDE beside every Δ.**

**5.6 Report dispersion and sign consistency, not just means** — per-seed Δ, sd(Δ), and sign agreement.

**5.7 Prefer fixed epoch budgets to early stopping** (arms differed by 35 epochs).

**5.8 Compute the seed budget before claiming; forbid Wilcoxon claims at n ≤ 3.** Required seeds: $n \approx (t \cdot sd(\Delta)/\Delta)^2$, here 17–113.

*(Protocol summary — the eight rules with their supporting evidence: Supplementary Table 11.)*

## 6. Discussion

### 6.1 What our results do and do not say

We do not claim STAL — or modules like it — is useless: Finding ⓪ shows the module moves mAP across a 5.03 pt range as σ varies. The claim is narrower: **at the sample sizes, calibers, budgets, and metrics the field actually uses, an effect of this size cannot be discriminated from seed noise.** "No measurable gain" and "no effect" are different statements; conflating them turns a null result into either false confidence or false dismissal. The stability across budgets follows: the obstacle is measurement resolution, which improves with seeds, not training length.

### 6.2 Implications for reporting practice

Three changes would make most failures detectable at submission, at near-zero cost: (1) **publish per-seed numbers** — three Δ values expose sd(Δ) and sign consistency that current papers aggregate away; (2) **fix and declare the caliber** — one sentence removes the 9.24 pt selection channel and the ρ sign-flip with it; (3) **report the MDE next to every Δ** — converting "we improved 0.4 pt" into "…where 3.3 pt is the smallest effect detectable," which is self-auditing. Reviewers should treat sub-point claims on three seeds as **incomplete rather than wrong**, and request the per-seed table.

### 6.3 Limitations

- **Two modules, one architecture family.** We audit two structurally different modules — STAL (assignment-side) and soft-clDice (loss-side) — on a single network family (YOLO26-seg), deliberately trading breadth for depth. The numerical results (sd(Δ), ρ, MDE) are properties of these pairings, not universal constants; the transferable claim is qualitative: at n = 3 these quantities are large enough to make sub-point claims undecidable, a claim that has survived both a module swap (§4.11) and a dataset swap (§4.10). Whether other module–network pairs behave the same is an open question our protocol lets others answer cheaply.
- **Two datasets, one model scale.** The paired comparison is replicated on CCU-Crack (§4.10), but both datasets are crack-specific and both use YOLO26m-seg; the numerical MDEs should not be assumed to transfer to other tasks or model sizes.
- **Domain metrics at two of three tiers** (50 and 300 ep).
- **Test-set size / IMR resolution**: 112 images, 9 crossing pairs (18 pooled with val) — a benchmark property, and why IMR is unusable for small effects.

- **One single-seed control** for the workers item (≈1 pt magnitude; rules out confounding, not a three-seed estimate).
- **A configuration discrepancy we disclose:** two of our IMR logs report 9 and 10 crossing pairs under the same touch tolerance — one enumerates ground-truth pairs (10, model-independent), the other counts pairs visible after prediction–GT matching inside the metric (9). All reported IMR numbers use the 9-pair configuration.
- **The 2 pt floor is domain-conventional**, calibrated to this task's sd(Δ) range; recalibrate per benchmark.

*(Budget-estimation lessons and Finding ⑦ seed counts: Supplementary Table 12.)*

## 7. Conclusion

We audited one representative crack-segmentation module across three budgets, three seeds, three calibers, and eight metrics, and replicated it on a second dataset and a second, structurally different module. The module has a genuine, parameter-dependent effect (⓪), yet yields no measurable gain: the paired correlation is caliber-dependent, flipping sign across calibers on identical runs and across datasets at matched budget and caliber (①, ⑧, §4.10); the caliber alone moves the paired spread 1.6× and SWA amplifies it 2.04× while reversing the sign (②); the effect's sign flips with budget while its magnitude stays below 0.62 pt on crack-seg and 0.85 pt on CCU-Crack (③); all seven domain metrics are undecidable, with IMR quantised to 11.1 pt and 20.0 pt (④); early stopping leaves arms 35 epochs apart (⑤); and decisively, a three-seed design cannot produce a nonparametric significance claim at any effect size (⑥), with the detectable effect 4.8–13.3× the observed one and 17–113 seeds needed to close the gap (⑦); a module-level control (soft-clDice, loss-side) is undecidable under the same protocol, its ceiling at 2.7× its observed effect (§4.11). Both replications reproduce the ceiling rather than the effect — the binding obstacle is inference, not this particular module or benchmark.

The contribution is not a module but a **minimum discernible protocol**: every rule is checkable in review from a single per-seed table, and none requires more compute than the community already spends.

## Data Availability

Two datasets were used. The primary benchmark, **crack-seg**, is an existing public dataset distributed by Ultralytics: its GitHub entry point — the dataset configuration page at https://github.com/ultralytics/ultralytics/blob/main/ultralytics/cfg/datasets/crack-seg.yaml — documents the dataset and links the ~91.6 MB archive used here (further documentation: https://docs.ultralytics.com/datasets/segment/crack-seg/), and its source is credited to Roboflow Universe (`crack-bphdr`, University, 2022, open source); the split used here is the one recorded in §3.1. The second dataset, **CCU-Crack**, was newly captured for this study; it is not publicly released, because the data-collection agreement does not permit it, and its imagery and labels are available from the corresponding author on reasonable request (ground-truth statistics in Supplementary Table 2). All data generated by this study — the training configurations and per-epoch validation curves of all 33 runs, the analysis code that regenerates every table and figure in this paper, the domain-metric definitions, the evaluation logs from which every domain-metric value was transcribed, and the verification scripts — are released with this article at https://doi.org/10.5281/zenodo.23373044 under AGPL-3.0. Model checkpoints are not shipped; each can be regenerated from its released configuration in ≈21 minutes on a 24 GB GPU.

## References

1. Perri, V., Cimichella, S., Crispino, M., & Toraldo, E. "A Comparative Investigation of YOLO26 and RF-DETR for Thin Crack Detection and Segmentation in UAV-Derived Airport Pavement Orthophotos." *Applied Sciences*, 16(17), 8608, 2026. DOI 10.3390/app16178608. — Source of the length-based **hallucination index (HI)** and **model error index (MEI)** calibers. The

remaining domain metrics (**FR**, **SLR**, **SLP**, **IMR**) are defined in this study and implemented in `crack_topology_metrics.py`, released with the code; they are **not** taken from a published benchmark, and we make no claim that they constitute a standard.

2. König, J., Jenkins, M., Mannion, M., Barrie, P., & Morison, G. "What's Cracking? A Review and Analysis of Deep Learning Methods for Structural Crack Segmentation, Detection and Quantification." arXiv:2202.03714, 2022.
3. Picard, D. "torch.manual_seed(3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision." arXiv:2109.08203, 2021.
4. Bouthillier, X., et al. "Accounting for Variance in Machine Learning Benchmarks." *Proceedings of Machine Learning and Systems (MLSys)*, 2021.
5. Åkesson, J., Töger, J., & Heiberg, E. "Random effects during training: Implications for deep learning-based medical image segmentation." *Computers in Biology and Medicine*, 180, 108944, 2024. DOI 10.1016/j.compbiomed.2024.108944. (*nnU-Net, 50 seeds; reports that paired t / Wilcoxon tests on seed-level differences are unreliable indicators of true algorithmic differences.*)
6. Redmon, J., et al. "You Only Look Once: Unified, Real-Time Object Detection." CVPR, 2016; Ultralytics YOLOv8/YOLO26 documentation, 2023–2026.
7. Zhang, S., et al. "Bridging the Gap Between Anchor-based and Anchor-free Detection via Adaptive Training Sample Selection (ATSS)." CVPR, 2020.
8. Ge, Z., et al. "OTA: Optimal Transport Assignment for Object Detection." CVPR, 2021.
9. Shit, S., Paetzold, J. C., Sekuboyina, A., et al. "clDice — A Novel Topology-Preserving Loss Function for Tubular Structure Segmentation." CVPR, 2021, pp. 16555–16564. DOI 10.1109/CVPR46437.2021.01629.
10. Izmailov, P., et al. "Averaging Weights Leads to Wider Optima and Better Generalization." UAI, 2018.

## Figure and Table Legends

**Figure 1** — Audit pipeline: from a single module to a discriminability verdict.

**Figure 2** — 40-ep tier per-epoch validation curves, 3 seeds x 2 arms.

**Table 1 — Base tier: STAL vs λ0 at the 40-ep budget (2 arms × 3 seeds, mask mAP50(M)).**

**Figure 3** — Sigma response; range 5.03 pt, all settings below baseline.

**Figure 4** — (a) 50-ep curves with the SWA window; (b) sd(Delta) by caliber.

**Figure 5** — Delta sign flip across 40/50/300 ep budgets.

**Figure 6** — Cross-dataset replication at the 40-ep budget under the win10 caliber. (a) Per-seed paired values; CCU-Crack is systematically easier than crack-seg. (b) Per-seed differences Δ scatter widely around the ±2 pt floor: individual seeds cross it on both datasets (−2.76 pt crack-seg, −2.19 pt CCU-Crack) though no *paired mean* reaches it. (c) MDE dwarfs the observed |mean Δ| by 8.5× (crack-seg) and 3.6× (CCU-Crack).

**Table 2 — Module-B control: soft-clDice (λ = 0.05) vs λ0, crack-seg, 2 arms × 3 seeds × 40 ep, mask mAP50(M).**

## Author contributions

The author conceived and designed the study, implemented the training and evaluation pipeline, ran all experiments, performed the analysis, and wrote the manuscript.

## Funding

This work was supported by the Chongqing Jiaotong University Postgraduate Joint Training Base (grant number XJLHPYJD202511). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## Additional Information

Competing interests: The author declares no competing financial and non-financial interests.

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.docx