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.

Cross-view agreement rescoring improves tiled small-object detection without retraining

Hyun-Woo Na
Hanbat National University
Eunkyung Kim
Hanbat National University
Haneol Jang
hejang@hanbat.ac.kr

Hanbat National University

Article

Keywords:

Posted Date: September 29th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10908559/v1

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

Additional Declarations: No competing interests reported.

Cross-view agreement rescoring improves tiled small-object detection without retraining

Hyun-Woo Na[1], **Eunkyung Kim**[2,*], **and Haneol Jang**[1,*]

[1]Department of Computer Engineering, Hanbat National University, Daejeon, Republic of Korea
[2]Hanbat National University, Department of Artificial Intelligence Software, 109 Jiphyeonbuk-ro, Sejong 30139, Republic of Korea
[*]Corresponding authors: ekim@hanbat.ac.kr; hejang@hanbat.ac.kr

ABSTRACT

Tiled inference is the standard route to small-object detection in large images, yet its merging stage compares confidence scores produced under different framings on a single axis — and average precision is a ranking metric. An oracle decomposition — replacing predicted scores with their ground-truth quality — shows that most of the recoverable headroom lies in score ranking rather than box selection. We show that the signal needed to repair the ranking is already computed and then discarded: the agreement of overlapping tiles. Normalising each detection's re-detection rate by its effective view set — only those views whose crop could actually have seen the object — yields a rescoring rule that requires no training, no labels, and no additional forward passes. On VisDrone test-dev with YOLOv8s it improves small-object average precision over slicing-aided hyper inference by $+1.05$ (95% confidence interval $+0.89$ to $+1.23$). A single frozen setting transfers unchanged to five benchmark splits and 110 detector–dataset configurations: across 1,095 pairings spanning five merge rules, rescoring improves average precision in 93%, including 96% on top of a recent crop-margin-aware merge, and held-out gains match validation. The added cost is CPU-only post-processing, about ten milliseconds per image at the median; the GPU cost is unchanged.

Introduction

A drone surveying a disaster zone for survivors, a conservation aircraft counting animals across a savanna, a satellite watching harbours and infrastructure — many of the most consequential uses of automated vision share one property: the objects that matter occupy a dozen pixels in an image of tens of millions. Detecting them reliably, without retraining a model for every new sensor, altitude or region, is a bottleneck across scientific and humanitarian applications, not only a computer-vision benchmark exercise. The standard remedy is tiled inference (Fig. 1a): cut the large image into overlapping crops, run the detector on each crop and on the full image, and merge the results. Slicing-Aided Hyper Inference (SAHI)[1] made this recipe the de facto baseline, and a large body of follow-up work has refined *where* to crop — adaptively sized slices[2,3], density-guided crops[4], attention- or content-gated crops[5,6] — or *how* to merge boxes, from Soft-NMS[7] and weighted box fusion[8] to crop-margin-aware suppression rules[4].

This paper argues that both threads overlook the pipeline's quietest defect: the *scores*. A tile magnifies its content, so the same object is seen at different effective resolutions by the full-image view and by each tile, and the confidence scores produced under these different framings are not mutually comparable. The merging stage nevertheless ranks them on a single axis — and average precision is a ranking metric, so mis-calibration between views converts directly into lost accuracy. An oracle decomposition (Fig. 2) makes this quantitative: given the boxes that tiled inference already produces, most of the recoverable headroom lies in re-ordering the scores, not in finding better boxes.

Our second observation is that the signal needed to repair the ranking is already computed and then thrown away. Overlapping tiles re-photograph the same object under different framings; a real object is re-detected by essentially every view that could have seen it, while background-texture false positives and boundary artefacts are not reproduced. Standard merging treats these duplicates as waste to be suppressed. We treat them as repeated measurements. The resulting rule, cross-view agreement rescoring (abbreviated CAR hereafter), multiplies each detection's score by a per-detection agreement factor before the merge (Fig. 1). The critical design choice is the denominator of that agreement: dividing by all views, as ensemble methods do[8], drives the method *below* the baseline, whereas dividing only by the views that could have seen the detection — the effective view set — reverses the sign. CAR needs no training, no labels, no architectural access, and adds zero forward passes; it is a CPU-only re-ranking layer inserted between detection and merging.

Consistency signals have been used elsewhere — for label-free accuracy estimation from pre-NMS prediction consistency[9,10], for boundary-restricted refinement[11], and for grouped-evidence fusion at the cost of extra inference[12] — and confidence calibration for detectors is an active field[13,14]. To our knowledge, however, no prior work uses visibility-normalised

Table 1. Main result: VisDrone test-dev (held-out; single evaluation run), YOLOv8s, identical views and forward passes throughout. The AP_s gain of CAR over SAHI+NMS is $+1.05$ (paired bootstrap 95% CI $[+0.89, +1.23]$, 10^4 resamples).

	AP	AP_{50}	AP_s
single-pass inference	16.02	27.33	7.08
tiled (SAHI + NMS)	20.83	36.25	13.53
tiled + CAR (ours)	**22.53**	**39.02**	**14.58**

cross-view agreement to repair the score *ranking* inside a tiled-inference pipeline at zero additional inference cost; classical calibration, which applies monotone per-set transforms, cannot re-order detections at all (Results, "Cross-view agreement repairs the ranking").

We validate CAR at, we believe, unusual breadth for a test-time method: one globally frozen hyper-parameter setting is evaluated on five benchmark splits (VisDrone[15], DOTA-v1.5[16], AI-TOD-v2[17,18]; validation and held-out test), 110 detector–dataset configurations spanning three detector families and two full capacity ladders[19,20], five merge rules, and open-vocabulary detectors[21,22]. Rescoring improves average precision in 93% of the 1,095 merge-rule pairings, including 96% on top of the 2025 crop-margin-aware CMA-NMS[4], and held-out gains match or exceed validation gains on both benchmarks with sealed test labels.

Results

The experiments answer five questions, in order. (i) *Where does tiled inference lose accuracy?* An oracle decomposition separates ranking from box selection (Fig. 2). (ii) *Does visibility-normalised agreement repair the loss?* The pre-registered main result on a sealed test split (Table 1), with the denominator isolated as the load-bearing design choice (Table 2) and labels shown to add almost nothing beyond it. (iii) *Is the repair orthogonal to the merge rule?* Five merge rules across the full grid (Table 3). (iv) *Does one frozen setting transfer?* Five benchmarks and 110 configurations including two held-out splits (Table 4), a failure-reversal case study (Table 4b) and the open-vocabulary regime, with qualitative evidence in Fig. 3. (v) *What does the gain cost, and where does it come from?* Timing across the grid and a per-class mechanism analysis (Fig. 4).

The bottleneck is ranking, not cropping

Before designing any method we measured what tiled inference *could* achieve if only its post-processing were perfect, holding the detector and the views fixed (Fig. 2). Given the pooled detections of a standard SAHI run, an oracle that replaces every confidence score by the detection's true overlap with its ground-truth object recovers the large majority of the achievable headroom: on VisDrone with YOLOv8s, 74% of the recoverable gap between the baseline and the box-selection ceiling is closed by re-ordering scores alone, and the same decomposition replicates almost unchanged on a second dataset and a second detector family (Supplementary Information). Fusing box *coordinates* (weighted box fusion[8]), by contrast, is actively harmful in this regime. The conclusion is that the boxes are already there; what tiled inference loses is their order. Any repair should therefore act on scores, before the merge — which is where CAR sits (Fig. 1).

Cross-view agreement repairs the ranking

CAR multiplies each detection's score s by $1 + \lambda \cdot \mathrm{agr}(d)$, where $\mathrm{agr}(d)$ is the fraction of the views that *could* have seen detection d (its effective view set $C(d)$; Methods) in which a matching same-class detection actually appears. Table 1 shows the main result on VisDrone test-dev — a split used exactly once, after every choice was frozen on the validation split. With YOLOv8s, tiling lifts single-pass inference from 16.02 to 20.83 AP; adding CAR lifts it further to 22.53 AP and from 13.53 to 14.58 AP_s ($+1.05$; paired bootstrap 95% CI $[+0.89, +1.23]$, 10^4 resamples).

The denominator of the agreement is the method's load-bearing part (Table 2). With identical detections, identical λ and an identical formula, normalising by the *total* number of views — the ensemble-fusion convention — lands 0.38 AP_s *below* the SAHI baseline, whereas normalising by the effective view set (Fig. 1b) lands 2.13 above it: a 2.51 AP_s swing from the denominator alone. Views that never had a chance to see a box must not vote against it. Excluding the detection's own view costs a further 0.34 AP_s, consistent with the own view being one more legitimate measurement.

A third control isolates ranking from selection, and is the most direct test of this paper's claim. Rescoring normally runs before the merge, so it changes both the order of the detections and which box survives each suppression cluster. Applying it *after* the merge instead freezes the surviving set — average recall is then numerically indistinguishable from the baseline (49.957 vs. 49.958 on RT-DETR-l) — and changes nothing but the order. That variant alone recovers 83–89% of the full gain across all four detectors we tested (RT-DETR-l 82.9%, RT-DETR-x 83.3%, YOLOv8s 89.1%, Faster R-CNN 83.7%;

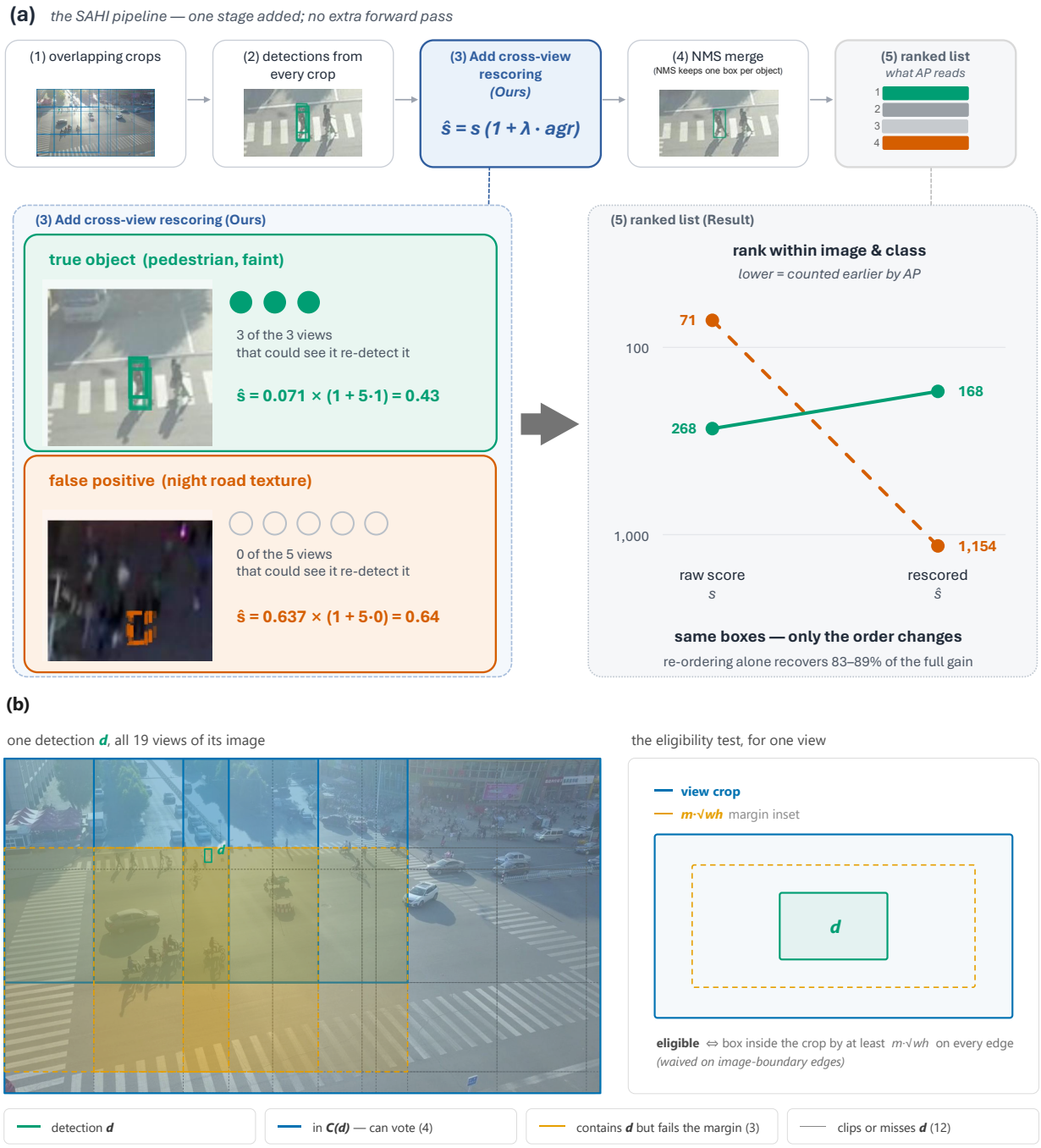


Figure 1. Where cross-view agreement rescoring sits in the tiled inference pipeline, and how it works. (**a**) A single rescoring stage — step (3) — is added to the SAHI pipeline (1)–(5); no forward pass is added and only the score entering the merge changes (+19 ms per image on the headline configuration). The left panel expands the added stage on two real VisDrone detections under the frozen setting (λ=5): a faint pedestrian re-detected by all three views that could see it, and a night-scene texture false positive re-detected by none of its five (night patch brightened 1.3× for legibility). The right panel expands the outcome in the unit the evaluator ranks in, an image–class pair (Methods): rescoring moves the object from rank 268 to 168 and the artefact from 71 to 1,154, while the surviving box set is unchanged. Across four detectors, re-ordering a fixed box set in this way recovers 83–89% of the method's full gain (Results). (**b**) The effective view set $C(d)$ on a real image, classified by the eligibility code of the released pipeline: views containing d with the required margin — inside the crop by $m\sqrt{wh}$ on every edge, waived at image-boundary edges — may vote (blue, 4); views that contain d but violate the margin (amber, 3) and views that clip or miss it (grey, 12) may not. Tints stack where crops overlap, so deeper colour marks regions covered by more views. Only crops that could have seen the box enter the denominator.

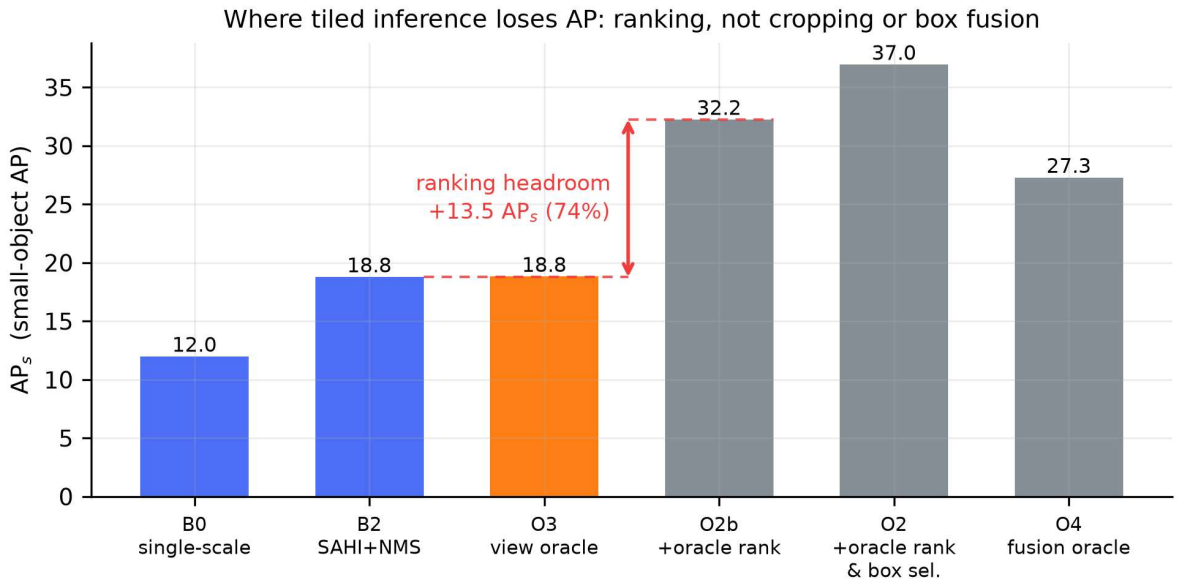


Figure 2. Oracle decomposition: tiled inference loses accuracy to ranking, not to cropping. Holding the detector, views and detections fixed, replacing only the confidence scores with an oracle closes 74% of the recoverable gap to the box-selection ceiling; selecting better boxes accounts for the remaining 26%. VisDrone, YOLOv8s, computed under the diagnostic probe protocol — a 274-image development half-split and a 640-px/0.25-overlap grid, run before the operating point was frozen (Supplementary Information) — so its absolute values are not directly comparable to Table 2, which uses the frozen 512/0.6 grid on the full validation split. Held-out test-dev numbers are in Table 1; the decomposition replicates on a second dataset and detector family at the frozen operating point (Supplementary Information).

Table 2. The denominator is the design choice that matters. Development split (VisDrone validation), YOLOv8s; identical detections, identical λ=5, only the denominator of the agreement varies. The swing between the effective-view-set form and the all-views form is +2.51 AP$_s$ (95% CI $[+2.08, +2.76]$, $p < 10^{-4}$; +3.52 AP with CI $[+3.09, +3.78]$), paired bootstrap as in Methods.

Score used for merging	AP	AP$_s$
raw scores (SAHI+NMS, no rescoring)	25.85	20.07
agr $= \lvert M\rvert/\lvert V\rvert$: divide by *all* views	25.22	19.69
agr $= \lvert M\rvert/\lvert C(d)\rvert$ excluding d's own view	28.31	21.86
agr $= \lvert M\rvert/\lvert C(d)\rvert$: effective view set (ours)	**28.74**	**22.20**

development split). The accuracy that tiled inference leaves on the table is therefore recovered by re-ordering the boxes it already has, not by keeping different ones.

Two further controls bound what is being learned. A recalibration layer trained on ground-truth labels of the target domain — the strongest score-repair a practitioner could buy — adds only +0.05 AP$_s$ over label-free CAR, and classical monotone calibration cannot change the ranking at all: the agreement signal, not supervision, carries the information. Component ablations (Supplementary Information) show the same picture from the other side: removing the agreement term collapses the gain, removing everything else leaves most of it intact.

The gain is orthogonal to the merge rule

If mis-ranked scores are the real defect, repairing them should help under *any* downstream merge rule — including rules specifically designed for tiled inference. Table 3 tests this across the full grid: 110 detector–dataset configurations × two view generators × five merge rules, 1,095 evaluated pairings in total. Rescoring improves AP in 93.2% of them (95% CI $[90.5, 95.6]$, cluster bootstrap resampling the 110 configurations to respect their shared detectors and datasets; median +0.41), and in 88.0% of the small-object metric.

Two entries deserve comment. First, the SAHI *default* merge (Greedy-NMM with an intersection-over-smaller criterion) is

Table 3. Adding CAR under five merge rules. Each row pools all 110 detector–dataset configurations and both view generators (219 pairings per rule after removing configurations where a dump is unavailable). CMA-NMS is the crop-margin-aware merge of DAHI[4] (2025), re-implemented from its Algorithm 2 with published defaults.

Merge rule	Median ΔAP	Improved	%
NMS	+0.75	213/219	97
Soft-NMS (linear)[7]	+0.64	210/219	96
Soft-NMS (Gaussian)[7]	+0.43	209/219	95
CMA-NMS[4]	+0.51	210/219	96
Greedy-NMM (SAHI default)[1]	+0.09	179/219	82

Table 4. One frozen setting transfers. (**a**) Five benchmark splits, SAHI+NMS → +CAR. "Held-out" splits were evaluated once, after all choices were frozen. Relative gain is the median of per-configuration ΔAP/AP. The AI-TOD-v2 validation row counts 24 rather than 25 configurations because the SAHI-view detection dump for YOLO-World is unavailable on that split; this is the one configuration of the 110 missing here, and the same one excluded from the 219 pairings per rule in Table 3. (**b**) Failure-reversal case study (AP; in-domain RT-DETR, held-out splits only): on VisDrone test-dev tiling *lowers* AP below single-pass inference — the only two such regressions in the grid — and CAR reverses both into the largest wins; on AI-TOD-v2 test, where tiling already helps the same architecture, CAR still adds on top.

(a)

Benchmark	Held-out	Configs	Median ΔAP	Rel. gain	Improved
VisDrone val		25	+2.49	+11.2%	25/25
VisDrone test-dev	✓	25	+1.51	+9.4%	25/25
DOTA-v1.5 val		10	+3.07	+6.9%	10/10
AI-TOD-v2 val		24	+0.41	+5.6%	24/24
AI-TOD-v2 test	✓	25	+0.27	+7.2%	25/25

(b)

Benchmark (held-out)	Detector	single-pass	tiled (SAHI+NMS)	tiled+CAR
VisDrone test-dev	RT-DETR-l	23.06	22.90 ↓	**25.06**
VisDrone test-dev	RT-DETR-x	24.10	23.70 ↓	**25.83**
AI-TOD-v2 test	RT-DETR-l	18.70	22.47	**23.10**
AI-TOD-v2 test	RT-DETR-x	18.75	21.89	**22.43**

itself the outlier: on dense small-object scenes it merges chains of neighbours into envelope boxes and falls by up to 6.1 AP_s *below* plain NMS (Supplementary Information) — a practical warning independent of CAR, since it is what SAHI ships with. CAR still helps in 82% of those pairings, but there is little left for a re-ranking to recover once boxes have been fused away. Second, CMA-NMS: the most recent, crop-aware merge rule behaves exactly like the others under rescoring (96% improved). Strikingly, once CAR is applied the median AP of NMS (20.63) and CMA-NMS (20.64) become indistinguishable, although the rules differ without it. Repairing the scores levels the differences between suppression geometries — the choice of merge rule becomes secondary once the ranking is right, which is the ranking account's own prediction.

One frozen setting transfers: five benchmarks, 110 configurations

Every number in this paper uses one global setting (λ=5, one matching threshold, one NMS threshold; Methods), tuned once on VisDrone validation with YOLOv8s and never re-fit. Table 4 summarises the transfer. On every benchmark, every configuration improves: median relative gains range from +5.6% to +11.2% of baseline AP across drone photography (VisDrone), oblique aerial mosaics (DOTA-v1.5, +3.07 median AP) and satellite tiny objects (AI-TOD-v2, mean object size 12.8 px).

The gain also does not fade as the detector improves. Sweeping both YOLO capacity ladders end to end — ten checkpoints spanning a 26× range in parameters, over which baseline AP rises from 22.2 to 32.4 — leaves ΔAP without a monotone trend in capacity (Spearman $\rho = +0.32$, two-sided $p = 0.37$, $n = 10$ checkpoints), while the small-object gain *increases* (Spearman $\rho = +0.71$, two-sided $p = 0.02$, $n = 10$). The reason is visible in the agreement itself: the mean agreement of the false positives sits at 0.334–0.347 for every one of the ten detectors, so the separation the rescoring exploits does not close as the model gets better.

The two held-out rows are the strongest evidence. On both benchmarks whose test labels were sealed until a single final

run, the *relative* gains match the validation gains: preserved on VisDrone (9.4% vs. 11.2%, on a larger baseline AP) and slightly exceeding validation on AI-TOD-v2 (7.2% vs. 5.6%) — nothing in the development numbers was an artefact of split choice. The grid also contains a case where tiling itself *fails* (Table 4b): for RT-DETR[20] trained in-domain on VisDrone, adding tiles lowers AP below single-pass inference. CAR does not just soften this failure — it reverses it, turning the only two non-degenerate regressions on this split into its largest wins. The regression is systematic rather than incidental: it appears in both capacities on both VisDrone splits, and a diagnosis on the development split locates its cause in what tiling costs rather than in the architecture. Counting the tile detections added per additional ground-truth object recovered gives 59 for YOLOv8s, 110 for Faster R-CNN, 415 and 437 for RT-DETR-l and -x — a perfectly monotone ordering with the sign and size of the tiling effect (Spearman $\rho = -1.00$; $n = 4$ detectors, reported descriptively). The same RT-DETR architecture gains from tiling when its full-image view is weaker (COCO weights +3.2, AI-TOD-v2 weights +4.0 AP), so the deciding factor is the ratio between what tiles add and what they flood, not the detector family. On AI-TOD-v2 test, where tiling already helps the same architecture, CAR still adds on top — the repair is not specific to the failure regime. Finally, the rule transfers to open-vocabulary detectors, whose text-similarity scores are notoriously mis-calibrated: YOLO-World[21] and Grounding DINO[22] both gain on VisDrone under the same frozen setting (Supplementary Information) — a regime where fine-tuning-based alternatives are inapplicable in principle, because the class set is unknown at training time. Fig. 3 shows representative qualitative cases at matched detection budgets.

Where the gain comes from

Because agr is computed from geometry alone, its mechanism can be read off directly (Fig. 4). Two effects dominate. For classes whose instances span multiple tiles, agreement concentrates score mass on consistently re-detected boxes and cleans boundary artefacts; the largest per-class win in DOTA (*harbor*, +14.5 AP) is exactly the class with the largest average tile span. For abundant tiny classes, the effect is budgeted recall: rescoring pushes true detections above a fixed per-image detection budget (top-K with K the number of annotations in that image, an analysis budget we impose, not the evaluator's), so more real objects survive at a matched budget. The distribution of gains across images is broad rather than driven by outliers (Supplementary Information).

What decides which classes gain, however, is not the true positives but the false positives. Across the ten VisDrone classes, the class-level gain tracks the *mean agreement of the false positives* more tightly than any other quantity we measured (Spearman −0.93, Pearson −0.98); the agreement of the true positives is nearly uninformative on its own. The cleanest illustration is a matched pair: *truck* and *awning-tricycle* have essentially the same true-positive agreement (0.707 vs. 0.701) yet gains that differ ninefold (+5.11 vs. +0.57 AP). What separates them is the false-positive side (0.337 vs. 0.454).

This has a concrete cause. Splitting false positives into those that overlap a ground-truth object of *another* class and those on pure background separates their agreement sharply (0.48–0.63 vs. 0.21–0.42): a confusion between neighbouring classes is a real object, so every view that can see it re-detects it, and no amount of cross-view evidence will suppress it. The weak classes are exactly the ones dominated by such confusions — *awning-tricycle* 41.4% (mostly onto *tricycle*), *bicycle* 31.4% (onto *motor*), *people* 28.3% (onto *pedestrian*) — against 11–12% for *car* and *pedestrian*. Background-texture false positives — road markings, building facets, wave crests — behave the opposite way: they are re-detected by almost no view that could have seen them, and fall in rank. The boundary condition for the method is therefore semantic rather than geometric: agreement removes the false positives that framing destroys, and cannot touch those that are objects in their own right.

Cost: CPU-only, bounded by detection density

CAR adds no forward passes; its cost is post-processing. On the headline configuration (VisDrone, YOLOv8s, 15.1 views per image on an NVIDIA A40), the pipeline spends 112.8 ms per image on GPU forward passes and 19.1 ms on merging in both variants; CAR adds 18.7 ms of CPU agreement computation, +14.2% of end-to-end latency (7.6→6.6 frames per second). Across all 26 timed benchmark–detector configurations the median overhead is +10.0 ms per image; it scales with detection density rather than image size, reaching +139.5 ms in the densest case (DOTA, 1,131 detections per image) — a regime where the tiled forward pass itself is an order of magnitude more expensive as well. The agreement stage parallelises trivially across images.

Discussion

Tiled inference research has concentrated on where to crop and how to merge. The oracle decomposition and the 1,095-pairing battery together indicate that a third axis — whether the scores entering the merge are comparable at all — carries most of the recoverable headroom and can be repaired for free, from information the pipeline already computes. Three implications follow.

First, the result re-orders practical priorities. A practitioner running SAHI today gains more from two zero-cost changes — switching the default Greedy-NMM merge to plain NMS, and rescoring with cross-view agreement — than from adopting a more sophisticated merge rule: the SAHI default falls by up to 6.1 AP_s below NMS on dense small objects, and once scores

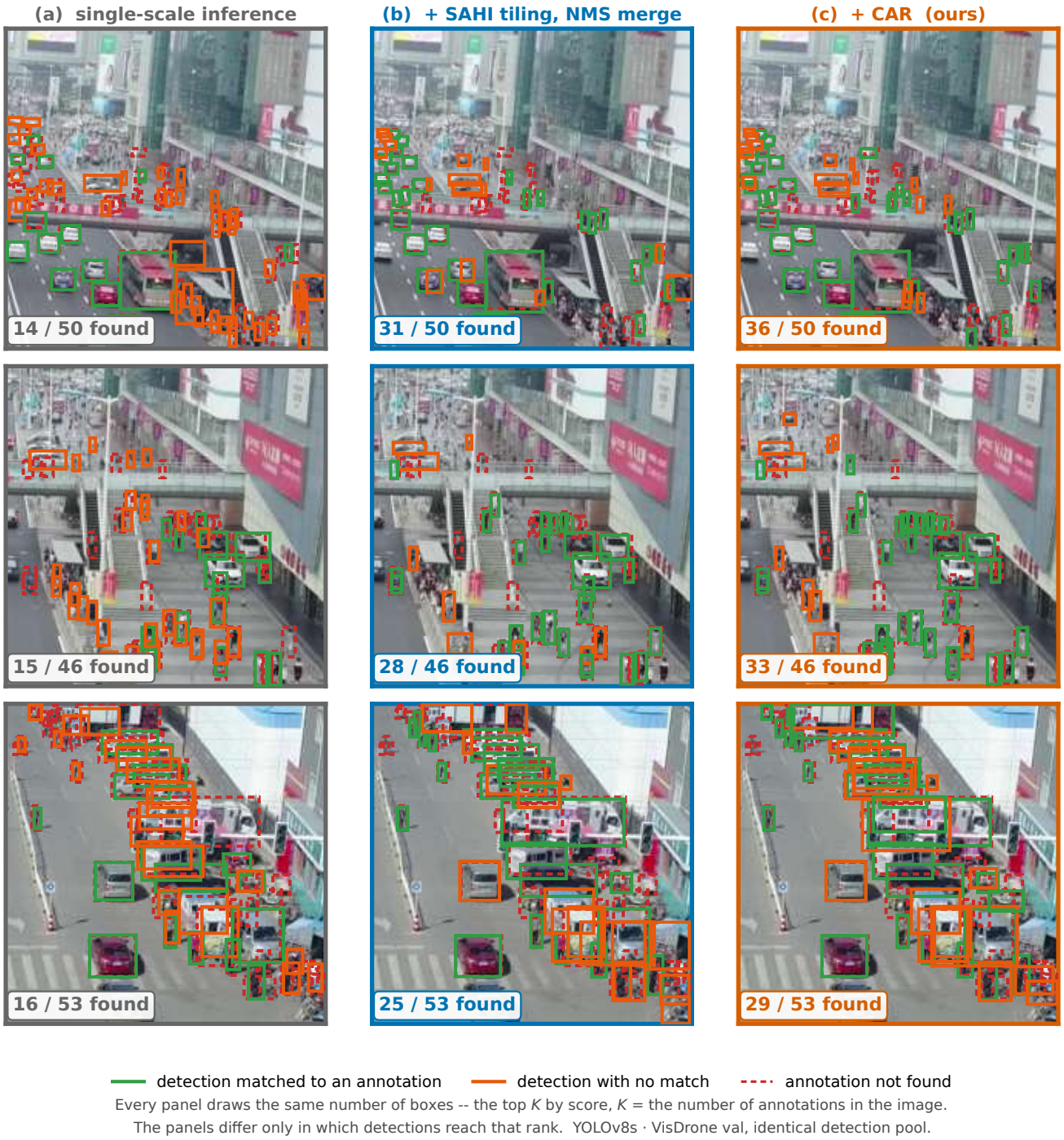


Figure 3. Qualitative effect at a matched detection budget. Three VisDrone validation scenes, one per row. (**a**) Single-scale inference. (**b**) SAHI+NMS: confident background-texture false positives (orange) crowd out low-scored real objects. (**c**) With CAR, consistently re-detected objects rise above the budget cut-off and single-view artefacts fall below it. Further examples, including the RT-DETR failure-reversal cases and held-out test-dev scenes, are in the Supplementary Information.

are repaired the remaining differences between suppression rules largely vanish (median AP 20.63 vs. 20.64 for NMS vs. CMA-NMS under CAR). Score repair, not suppression geometry, is where the leverage is.

Second, the breadth of the transfer — one frozen scalar across five benchmark splits, three detector families, two capacity ladders, five merge rules and the open-vocabulary regime, with held-out relative gains matching validation — supports the

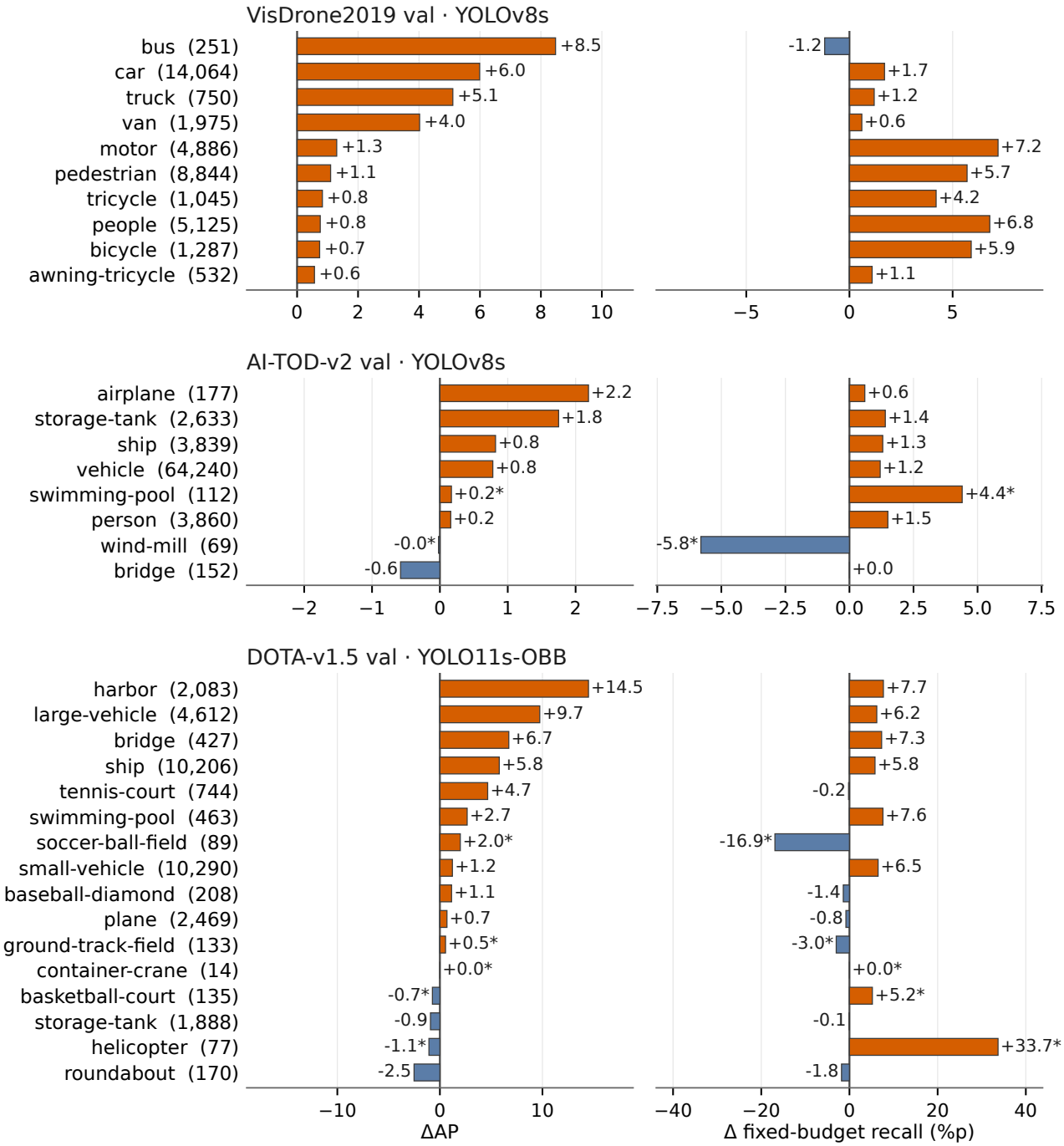


Figure 4. Which classes benefit, and why. Per-class change in AP (left) and in fixed-budget recall (right) when CAR is added on identical detections, one row per benchmark; orange bars are gains, blue are losses, and axis ranges differ between benchmarks. Numbers in parentheses are ground-truth instance counts; classes with fewer than 150 instances are marked ∗. Classes whose instances span multiple tiles (e.g. *harbor* in DOTA-v1.5) gain most from agreement-based cleaning of boundary artefacts; abundant tiny classes gain budgeted recall as true detections are pushed above the per-image detection cut-off (top-K per image with K the number of annotations; Methods).

reading that mis-ranking between views is a *structural* property of tiled inference, not a quirk of one detector or dataset. The mechanism requires only that real objects be re-detected more consistently than artefacts across overlapping framings; nothing in it is specific to aerial imagery.

Third, the agreement signal is related to, but distinct from, two literatures. Prediction-consistency signals have been used to *estimate* detector accuracy without labels[9,10]; we use a visibility-normalised variant to *improve* accuracy, inside the pipeline, at ranking granularity. Confidence calibration[13,14] adjusts probabilities via monotone maps that preserve per-set order and therefore cannot recover ranking-induced AP loss; non-monotone *learned* rescorers can[23], but need labelled training data, and our label-trained control shows that such a recalibrator adds almost nothing (+0.05 AP_s) beyond the label-free agreement signal.

Scope CAR repairs the *order* of detections. It is not a probability calibrator: the multiplier is applied and the scores are then rescaled to $[0,1]$, an order-preserving transform, so we make no claim about expected calibration error or any other reliability metric. Where we compare against calibration methods, the comparison is about what each can do to ranking.

Limitations A post-hoc diagnosis of the per-class results, carried out after the setting was frozen, identified three regimes where the mechanism weakens; none is repaired here, because changing the rule after seeing these results would forfeit the pre-registration the main claim rests on.

First, the agreement match is a *localisation* test, and at very small sizes it becomes a sub-pixel one. Requiring IoU $\geq \tau_a$ between two views of a box with short side s permits a centre displacement of at most $s/7$; for AI-TOD-v2's *person* (median short side 6 px) that is 0.86 px, while single views already localise to about a pixel. Decomposing the mismatches for that class, 84% are boxes that were found but fell short of τ_a rather than boxes that were missed — the agreement is measuring regression jitter, not existence. Relaxing τ_a does not repair this: false positives are recovered at nearly the same rate as true ones, and the frozen 0.75 maximises average precision over the values we tried. This, and not a shortage of views, explains AI-TOD-v2's small gains (its median $|C(d)|$ is 4–5, and fewer than 1% of detections have $|C(d)| \leq 1$).

Second, the eligibility rule admits every view that geometrically contains the box, including the full-image view, whose scale factor on DOTA-v1.5 is only 0.16: a 30 px tank enters the network at under 5 px there, and the agreement concentrates near $1 - 1/|C(d)|$ on that benchmark. One might expect that vote to be uninformative and worth discarding. It is not. Removing the full-image view from the denominator costs 2.4 AP on DOTA-v1.5 and 1.2 AP on VisDrone, and helps only 8 of 33 classes (of the 34 classes in Fig. 4, DOTA's *container-crane* is excluded as a tie: this checkpoint never detects it, so its AP is zero under every variant); measured before merging, the full view is in fact more likely to agree with a true positive than with a false one in every class we examined. We found no view-selection variant that improves on the rule as defined.

Third, and most fundamentally, agreement can only remove false positives that framing destroys. A detection that overlaps a real object of a neighbouring class, or a background structure that is genuinely circular or linear, is re-detected by every view that can see it, so its agreement is as high as a true positive's — the failure mode analysed above for *awning-tricycle* and *bicycle*. Consistently with this, one class in the whole grid loses accuracy for reasons internal to the method.

The overhead grows with detection density and reached +139.5 ms per image in our densest configuration; applications with hard latency budgets should account for it, although the same configurations also pay proportionally more for the tiled forward passes themselves. Open-vocabulary transfer is positive but uneven: on AI-TOD-v2 the open-vocabulary detector is effectively blind (baseline AP near zero), and no rescoring can recover detections that were never made. Finally, CAR re-ranks existing boxes; it cannot create recall beyond the view set, and it inherits whatever biases the frozen detector has.

In sum, tiled inference loses accuracy to ranking, not to cropping — and the duplicates it discards are precisely the measurements needed to repair the loss. We expect visibility-normalised agreement to be useful wherever a detector sees the same scene more than once.

Methods

Views and notation

Given an image, the view set V consists of the full image resized to the detector's input resolution plus an overlapping SAHI grid[1] of square tiles (side S=512, overlap ratio r=0.6). The frozen detector runs once per view; detections are mapped back to original-image coordinates, giving one pooled set per image with a score floor of 0.02. Nothing upstream of rescoring differs from a standard SAHI run, and the number of forward passes is identical with and without CAR. The second view generator used in the merge-rule battery is a content-independent re-implementation of adaptive slicing[2]: the tile count is fixed by resolution (6 tiles if $\max(W,H) < 1{,}600$ px, else 12), split into a grid matching the image aspect ratio, with tile size back-computed and overlap ratio 0.25, plus the full-image view. CAR itself is identical under both generators. Both external baselines were re-implemented from the descriptions in their papers — ASAHI from its resolution-based slicing rule, CMA-NMS from DAHI's Algorithm 2 with published defaults — as we could not locate official public implementations at the time of writing; both re-implementations ship with the released code.

Effective view set and agreement

Fig. 1b illustrates the eligibility rule on a real image, classified by the released code itself. For a detection d with box $b(d)$, the effective view set $C(d)$ contains the views in which $b(d)$ lies fully inside the view's crop with a margin of m=0.5 times the box scale — defined as $\sqrt{wh}$ of the box — from every crop edge; the margin is waived on any edge where the crop edge coincides with the image boundary, since an object there cannot be truncated by the crop. $C(d)$ formalises "the views that could have seen d": views whose crop excludes or clips the box are not eligible to vote. The matched subset $M(d) \subseteq C(d)$ contains the eligible views that contribute *at least one* same-class detection whose intersection-over-union with $b(d)$ is at least τ_a=0.75 — matching tests existence per view, not a one-to-one assignment, and d's own view counts when it is eligible. Matching is class-conditional throughout; for open-vocabulary detectors the prompt-to-class mapping defines the classes, so agreement is computed per prompt-mapped class. The agreement is $\mathrm{agr}(d) = |M(d)|/|C(d)|$, and the rescored confidence is

$$\hat{s}(d) \;=\; s(d)\,\big(1+\lambda\cdot\mathrm{agr}(d)\big), \qquad \lambda = 5.0, \tag{1}$$

The two overlap thresholds in the pipeline serve different purposes and act at different stages: the agreement match (τ_a) only asks whether a view re-detects d and deletes nothing, whereas the final NMS threshold (0.6) removes boxes. Rescoring changes what NMS keeps, but the two tests are never applied to the same decision. The rescoring is followed by a rescale of all scores to $[0,1]$, which leaves the ranking unchanged. We use the linear form because it is the simplest monotone family that preserves score order within equal-agreement groups; the saturation of the gain in λ (Supplementary Information) indicates the exact functional form is not critical, and richer forms are left to future work. Rescoring runs *before* the merge, so it changes both the final ranking and which box survives each suppression cluster. Class-wise NMS with an intersection-over-union threshold of 0.6 produces the final set. All values above were fixed once on VisDrone validation with YOLOv8s and reused verbatim everywhere, including all other detectors, benchmarks, merge-rule experiments and the open-vocabulary runs. None of them is delicate: on the development split, every $\lambda \in [2,10]$ lies within 0.15 AP of the best value (λ=1 already recovers 89% of the full gain), and varying τ_a over $[0.6,0.8]$ or m over $[0.25,0.75]$ moves AP by at most 0.33 (Supplementary Information).

Benchmarks, detectors and evaluation

Absolute AP values reported here are not comparable to those in the original SAHI, ASAHI or DAHI papers, which use different detectors, slicing-aided fine-tuning and partly different metrics (size-wise AP_{50}); all comparisons in this paper are paired, within-pipeline contrasts on identical detection pools. We evaluate on VisDrone2019-DET[15] (validation, 548 images; test-dev, 1,610 images), DOTA-v1.5[16] (validation), and AI-TOD-v2[17,18] (validation, and the 14,018-image test split with 376,121 instances). Detectors span the YOLOv8 and YOLO11 capacity ladders (n–x)[19], RT-DETR-l/x[20], Faster R-CNN[24], and the open-vocabulary YOLO-World[21] and Grounding DINO[22], in in-domain and cross-domain (COCO-trained[25], forced class mapping) variants — 110 detector–dataset configurations in total. Checkpoints are public or, where no public in-domain checkpoint exists (RT-DETR and Faster R-CNN on VisDrone and all AI-TOD-v2 detectors), trained once by us with standard recipes and then frozen; training details are in the Supplementary Information, and these models are evaluation subjects, not part of the method. Accuracy is computed with the official COCO evaluator[25] per benchmark protocol (`maxDets` 500 for VisDrone, 2,000 for DOTA-v1.5 — which has up to 1,814 annotations in one image — and 1,500 for AI-TOD-v2; the evaluator ranks and truncates detections *per image and category*, so that pair is the unit in which ranking matters. Ranks quoted in the text and figures are positions in the full per-image–class list, before that truncation). DOTA detectors emit oriented boxes; these are converted to horizontal bounding boxes at dump time, and agreement matching and evaluation are both axis-aligned — CAR on oriented geometry is untested and left to future work. Because size-band names collide across benchmarks, small-object columns report COCO/VisDrone AP_s (area $< 32^2$) and, for AI-TOD-v2, its very-tiny band AP_{vt} (2–8 px), never mixed in one column. The per-class analysis additionally uses a fixed-budget recall: for each image we keep the top K detections by score across all classes, with K the number of annotations in that image. Whenever the detector supplies at least K boxes the number of predictions equals the number of targets, so precision, recall and F_1 coincide at that operating point; this is our own analysis budget and is unrelated to the evaluator's `maxDets`. The label-trained recalibration control referenced in Results is a small gradient-boosted rescorer ($\leq$ 100 trees over 13 geometry and score features) trained with target-domain ground truth; its full specification, an isotonic-regression variant and both label-use modes are in the Supplementary Information.

Statistics

The pre-registered primary endpoint — CAR versus SAHI+NMS on VisDrone test-dev AP_s — uses an image-level paired bootstrap with 10^4 resamples (fixed seed)[26]. All tests are two-sided at $\alpha = 0.05$; secondary metrics use Holm correction[27]. Reported bootstrap p-values are bounded below by the bootstrap resolution, so we write $p < 10^{-4}$ rather than an exact value. Development and test splits were separated at the image level; each held-out split (VisDrone test-dev, AI-TOD-v2 test) was evaluated in a single run after all choices were frozen, enforced by a marker file recording timestamp and configuration.

Timing

Forward-pass time was measured on an NVIDIA A40 (batch of views, 100 images, warm start). Post-processing stages (agreement, rescoring, merge) were timed on a 32-core CPU host over 120 sampled images per configuration, median of three repeats, single worker per image; the sweep covers 26 benchmark–detector configurations. Forward cost is identical in baseline and CAR by construction.

Data availability

The datasets analysed in this study are publicly available from their original distributors under their respective academic licences: VisDrone2019-DET, DOTA-v1.5 and AI-TOD-v2. We do not redistribute image data. The full 110-configuration result grid, the per-detection dumps and the numbers behind every table and figure are deposited on Zenodo (https://doi.org/10.5281/zenodo.22256020; published 6 September 2026, open access; data under CC BY 4.0, code under AGPL-3.0) and are additionally available from the corresponding authors.

Code availability

The full experimental pipeline (view generation, agreement computation, rescoring, evaluation and table generation) is released under the AGPL-3.0 licence; an anonymised repository is made available to the reviewers during peer review, and the archived version is deposited with the Zenodo record above. Third-party components retain their own licences; pre-trained checkpoints are used unmodified from their public sources.

References

1. Akyon, F. C., Onur Altinuc, S. & Temizel, A. Slicing aided hyper inference and fine-tuning for small object detection. In *IEEE International Conference on Image Processing (ICIP)*, 966–970 (2022).
2. Zhang, H., Hao, C., Song, W., Jiang, B. & Li, B. Adaptive slicing-aided hyper inference for small object detection in high-resolution remote sensing images. *Remote. Sens.* **15**, 1249 (2023).
3. Moretti, F., Jin, Y. & Mario, G. Adaptive slicing-assisted hyper inference for enhanced small object detection in high-resolution imagery (2026). ArXiv:2604.19233, 2604.19233.
4. Suárez-Ramírez, J., Santana-Cedrés, D. & Monzón, N. DAHI: a fast and efficient density aided hyper inference technique for large scene object detection. *Pattern Recognit.* **171**, 112228 (2025). DOI 10.1016/j.patcog.2025.112228.
5. Wang, H., Li, H., Chen, W. & Zhang, X. ViCrop-Det: Spatial attention entropy guided cropping for training-free small-object detection. *arXiv preprint arXiv:2604.26806* (2026).
6. Riyadh, R., Khan, A. U., Gohar, I. & Behzad, M. ROI-Gated SAHI: Content-adaptive slicing-based inference for efficient object detection (2026). ArXiv:2608.23923, 2608.23923.
7. Bodla, N., Singh, B., Chellappa, R. & Davis, L. S. Soft-NMS – improving object detection with one line of code. In *IEEE International Conference on Computer Vision (ICCV)*, 5561–5569 (2017).
8. Solovyev, R., Wang, W. & Gabruseva, T. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image Vis. Comput.* **107**, 104117 (2021).
9. Yoo, S., Kwon, H., Hwang, J.-W. & Lee, K. Automated model evaluation for object detection via prediction consistency and reliability. In *IEEE/CVF International Conference on Computer Vision (ICCV)* (2025).
10. Manoharan, A., Yin, X., Helm, D. & Cheng, C.-H. Cumulative consensus score: Label-free and model-agnostic evaluation of object detectors in deployment. *arXiv preprint arXiv:2509.12871* (2025).
11. Shalmani, M. A., Boroujeni, A. R., Amini, A. & Yuan, J. S. From full boards to tiny defects: Scale-aware tile inference with topology-aware merging for high-resolution PCB defect detection. *arXiv preprint arXiv:2605.24726* (2026).
12. Xiao, Y. Group evidence matters: Tiling-based semantic gating for dense object detection. *arXiv preprint arXiv:2509.10779* (2025).
13. Kuzucu, S., Oksuz, K., Sadeghi, J. & Dokania, P. K. On calibration of object detectors: Pitfalls, evaluation and baselines. In *Computer Vision – ECCV 2024*, vol. 15079 of *Lecture Notes in Computer Science*, 185–204 (Springer, 2024). DOI 10.1007/978-3-031-72664-4_11.
14. Gilg, J., Teepe, T., Herzog, F., Wolters, P. & Rigoll, G. Do we still need non-maximum suppression? accurate confidence estimates and implicit duplication modeling with iou-aware calibration. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 4838–4847 (2024). URL https://arxiv.org/abs/2309.03110.

15. Zhu, P. *et al.* Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis Mach. Intell.* **44**, 7380–7399 (2021).

16. Xia, G.-S. *et al.* DOTA: A large-scale dataset for object detection in aerial images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3974–3983 (2018).

17. Xu, C. *et al.* Detecting tiny objects in aerial images: A normalized wasserstein distance and a new benchmark. *ISPRS J. Photogramm. Remote. Sens.* **190**, 79–93 (2022). URL https://arxiv.org/abs/2206.13996.

18. Wang, J., Yang, W., Guo, H., Zhang, R. & Xia, G.-S. Tiny object detection in aerial images. In *25th International Conference on Pattern Recognition (ICPR)*, 3791–3798 (2020). DOI 10.1109/ICPR48806.2021.9413340.

19. Jocher, G., Chaurasia, A. & Qiu, J. Ultralytics YOLOv8 and YOLO11. https://github.com/ultralytics/ultralytics (2023).

20. Zhao, Y. *et al.* DETRs beat YOLOs on real-time object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16965–16974 (2024).

21. Cheng, T. *et al.* YOLO-World: Real-time open-vocabulary object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16901–16911 (2024).

22. Liu, S. *et al.* Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision (ECCV)*, 38–55 (2024).

23. Pato, L. V., Negrinho, R. & Aguiar, P. M. Q. Seeing without looking: Contextual rescoring of object detections for AP maximization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020).

24. Ren, S., He, K., Girshick, R. & Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 91–99 (2015).

25. Lin, T.-Y. *et al.* Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 740–755 (2014).

26. Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap* (Chapman & Hall, 1993).

27. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6**, 65–70 (1979).

Author contributions

Conceptualization and methodology, H.-W.N. and H.J.; software, data curation and visualization, H.-W.N.; validation, formal analysis and investigation, H.-W.N. and E.K.; writing—original draft preparation, H.-W.N.; writing—review and editing, E.K. and H.J.; supervision, project administration and funding acquisition, E.K. and H.J. All authors have read and agreed to the published version of the manuscript.

Competing interests

The authors declare no competing interests.

Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2026-25490089).

Supplementary Files

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

- supplementary.pdf