Boxes to Scores: Distilling Box-Supervised Defect Detectors into Compact Image-Level Anomaly Scouts

Rosepetal

Abstract

Industrial inspection lines ultimately need one image-level decision — pass or flag — yet the most accurate defect models are box-supervised detectors: annotation-hungry to train and larger than the deployment slot they must fill. We ask whether the knowledge inside such a detector can be transferred *across tasks*: from box-level detection into a compact student that outputs only an image-level anomaly score and runs without its teacher. Our recipe converts cached detector predictions into pseudo-box targets, assigns them to the student's queries by Hungarian matching, and trains a small multi-scale transformer student jointly on these targets and the images' binary labels. In a controlled comparison, the distilled student outperforms the identical architecture trained on the same pool's image labels alone on both measured control seeds, by margins well beyond a pre-registered indistinguishability bar — more image-label supervision does not close the gap — and reaches 0.9614 AUROC against its strongest teacher's 0.9977 at a fraction of the parameter count. Box-supervised knowledge is transferable into image-level anomaly scoring; the detector's box-level targets, not extra image labels, carry the gain.

1. Introduction

An industrial inspection line renders one decision per image — pass or flag — yet the models that find defects most accurately reason per box: detection transformers trained on localized box annotations (Robinson et al. 2026; Lv et al. 2024). The asymmetry is paid twice: box annotations are among the costliest labels a line produces, and the detectors they buy are typically larger than the front-line scoring slot they would fill. A recent practitioner-driven critique of the field contends that a deployment's first need is a reliable image-level decision, not another leaderboard detector (Baitieva et al. 2025). The dominant OK-only alternative learns normality from defect-free images (Roth et al. 2022); its distillation-based line scores anomalies by student–teacher discrepancy against a task-agnostic frozen feature network mimicked on normal images (Batzner, Heckler, and König 2024) — a formulation that caps what distillation can transfer: a teacher that has never seen a defect cannot pass on what box-supervised detectors know about real defect appearance.

We ask whether that knowledge can transfer *across tasks*: from box-level detection into a compact student whose only deployable output is an image-level anomaly score, run without its teacher. The recipe caches trained detectors' predictions over the training pool, converts them into pseudo-box targets assigned to the queries of a small multi-scale transformer student by Hungarian matching, and trains the student jointly on those targets and the images' binary labels, exactly as implemented (Section 3).

This proposal has an obvious deflationary reading, and the paper is built to test it: boxes trivially imply image-level labels, so the same pool supports a much simpler recipe — the same architecture trained as a plain image-label classifier, no distillation machinery. We run exactly that control on a standard industrial anomaly benchmark (Bergmann et al. 2019); it loses on both control seeds measured, by margins of five and about three times the pre-registered bar (Section 5). Four contributions follow.

We present a cross-task box-to-score distillation recipe, described as-implemented. Cached detector predictions become Hungarian-matched pseudo-box targets for a compact multi-scale query-based student whose deployed output is a single, teacher-free image-level score; the description commits to the implemented mechanism — hard pseudo-box targets, and a first-seen-wins merge that leaves one teacher the sole effective supervision source — so the recipe is exactly reproducible and exactly falsifiable (Section 3).

We attribute the gain to the box targets through the paper's central deflationary control. The same architecture trained on the same pool's image labels, boxes withheld, under its original shorter recipe rather than a step-matched one, trails the distilled student on both measured control seeds — by 0.050 and 0.029 AUROC, five and about three times the pre-registered bar — and lands in the band of a

weak-label scout trained on far fewer labels: the image-label signal saturates, and the box targets carry the gain (Sections 4, 5, 7).

We measure a retention–compression trade-off. The distilled student retains 96.4% of its strongest teacher's AUROC at a fraction of the parameter count, and scores above the second detector teacher and the strongest OK-only baseline on the same held-out surface; parameter count is the only efficiency claim made — latency is unprofiled (Sections 5, 7, 9).

We report historical multi-surface consistency as secondary evidence. On four additional benchmark surfaces measured before the current single-benchmark freeze, the recipe scores in a band Section 6 reports cell by cell — each under a machine-checked train/eval-disjointness verdict, one with a disclosed score-polarity caveat — consistent with the mechanism not being an artefact of one dataset (Sections 6 and 9).

The evidence has disclosed boundaries — a single-seed distilled student compared against two measured control seeds (a third was deferred), a control kept on its original, shorter recipe rather than a step-matched one, an effectively single-teacher merge, no localisation claim, and no latency measurement — each stated where the affected claim appears and concentrated in Section 9. Section 2 formalises the setting, Section 3 the recipe as implemented, Sections 4–5 the control and results, Section 6 the historical cells, and Sections 7–10 analysis, positioning, limitations, and conclusion.

2. Problem setting

The supervision spectrum. Industrial anomaly-detection methods are ordered by the labels they consume. At the cheap end, OK-only methods model normal appearance from defect-free images and score deviation from it (Roth et al. 2022; You et al. 2022); image-level supervision adds one binary normal/defective label per image; box- and mask-level supervision requires per-defect spatial annotation, whose cost dominates dataset construction in practice (Božič, Tabernik, and Skočaj 2021; Cheng et al. 2025). Accuracy generally rises along this axis — and so does the annotation bill.

Boxes to train, scores to deploy. The setting studied here is deliberately asymmetric in time: box supervision is available while the student is trained, but nothing box-shaped survives into deployment. At inference the model receives one image and returns one scalar anomaly score; no teacher, cached prediction, or box pipeline is present. The expensive supervision is spent once, offline; the deployed interface is the cheapest one on the spectrum.

Preconditions. We state the setting's precondition upfront rather than in passing: the recipe presupposes that a box-annotated training pool and a defect detector trained on it already exist upstream. This is a scoped assumption, not a hidden one — annotation cost is moved into the teacher, not eliminated — and it matches a common industrial situation in which a capable detector exists but the deployment slot calls for a much smaller model emitting only a per-image decision. Conditional on that setting, the question this paper answers is whether the detector's box-level knowledge can be transferred into the image-level scorer.

Task definition. The student solves image-level anomaly scoring: map an input image x to a scalar anomaly score $s(x)$, higher meaning more likely defective, such that defective images rank above normal ones. The deployed output is this score and nothing else, and it is produced teacher-free — the detector participates in training only.

The scout. We call such a model a *scout*: a compact, standalone, front-line image-level scorer. The word names a deployment role, not a pipeline stage — no cascade, router, early exit, or downstream inspection machinery is assumed, built, or evaluated in this paper; a scout here is a complete deployable on its own.

Metrics. Headline metrics are threshold-free: image-level AUROC, with AUPR alongside, compares scoring quality independently of any operating point. False-alarm and missed-defect rates appear only descriptively, at a decision threshold frozen in advance and never tuned: no threshold is tuned on validation data, and no claim in this paper rests on any thresholded metric.

Non-goals. The student is trained with box targets and therefore emits boxes as a by-product, but we make no localisation claim: the deployed output is the image score alone, the emitted boxes are not a localisation capability, and their behaviour is examined only as a disclosed limitation (Section 7, Section 9).

3. Method: from boxes to scores, as implemented

The recipe has three ingredients: box-supervised detector teachers whose cached predictions become per-image pseudo-box targets; Hungarian matching that assigns those targets to the queries of a compact multi-scale student; and a decoupled image head, trained jointly with the matched-query losses, whose sigmoid is the only output the deployed scout emits. Figure 1 shows the pipeline. Everything below describes the implementation actually measured in Section 5, including two facts a cleaner idealisation would hide: the targets are hard pseudo-boxes, and the teacher merge leaves one teacher as the sole effective supervision source.

3.1 Teacher-target construction

The two box-supervised detector teachers pre-date this study: RF-DETR-S (Robinson et al. 2026), a 31.8 M-parameter detection transformer over a DINOv2-S backbone (Oquab et al. 2023), and RT-DETRv2 (Lv et al. 2024), a 20.5 M-parameter real-time detection transformer over a ResNet-18 backbone (vd variant) (He et al. 2016); every parameter count in this paper — teachers and student alike — is the sum of the model's checkpoint tensor elements, buffers included. The checkpoints that generate the distillation targets were trained not on the student's training pool but on the dataset's broader box-annotated detection split, whose overlap with the evaluation surface is disclosed and quantified in Section 4; the teacher rows reported in Section 5 come from separate, later checkpoints trained on the pool itself. Each teacher's predictions over the 3,750-image training pool are cached once and reshaped into per-image target records — predicted boxes with confidence scores — after which the teachers are never queried again: distillation runs entirely from the cached records, and no teacher is loaded at student training or inference time. One structural property of these targets matters throughout: detector teachers emit boxes on *every* image, defective and normal alike, so the target signal is dense precisely where ground-truth box supervision would be background-only — normal images receive low-confidence teacher boxes rather than empty target sets.

3.2 Target merge and Hungarian matching

The two teachers' target files are merged per image by keeping the first teacher's record wherever both provide one. Because the first teacher's records cover every image in the training pool, the second teacher contributes no effective supervision in the reported runs: the recipe, as implemented and as measured, is effectively single-teacher, with the stronger detector as the sole source of targets. We state this plainly rather than idealising the merge as an ensemble; per-teacher and true-union ablations are future work (Section 9).

Each image's targets are assigned to the student's 100 queries by exact linear-sum assignment — the Hungarian matching of set-based detection (Carion et al. 2020) — under the pairwise cost $-\mathrm{IoU} + 0.5\,L_1 - 0.5\,c$, where L_1 is the box-coordinate distance and c the student query's predicted confidence. Matched queries receive a binary classification loss toward their target plus box-regression losses, L_1 with weight 5.0 and GIoU with weight 2.0; unmatched queries are pushed toward background. The classification targets are hard: teacher confidences are stored in the target records but not propagated into the loss, so the algorithm as implemented is Hungarian-matched pseudo-box supervision rather than soft-logit distillation — Section 7 concedes this mechanical reduction explicitly.

3.3 Student architecture

The student backbone is a truncated ResNet-18 (He et al. 2016), kept through its third stage with the final stage dropped, yielding three feature scales at strides 4/8/16 with 64/128/256 channels. Per-scale 1×1 projections map every scale to a shared width $d = 256$, producing 5,376 tokens at the 256 px input resolution. A 2-layer pre-norm transformer encoder processes the full token set, and a 2-layer decoder attends over 100

learnable queries. Two heads sit on the decoder — a per-query binary logit, initialised with a 0.01-prior bias, and a 3-layer box MLP — and both exist to consume the matched-target losses: they are training scaffolding, not deployed outputs.

The deployed output comes instead from a *decoupled image head*: a LayerNorm-MLP over the mean-pooled encoder memory, whose sigmoid is the image-level anomaly score. Decoupling means the score is not a maximum over query confidences but a separately learned global readout of the encoder representation. The full student totals 6.81 M parameters, counted as the sum of checkpoint tensor elements with buffers included. At inference the student runs teacher-free: one forward pass, one scalar score.

3.4 Training objective

The student trains for 30,000 steps at batch size 16 with AdamW (learning rate 1e-4, weight decay 1e-4), in fp32, on 256 px letterboxed inputs, seed 1337. The total loss is the sum of the matched-query terms of Section 3.2 and an image-level binary cross-entropy (weight 1.0) between the image head and the image's binary normal/defective label. This image-label term is deliberately shared with the control of Section 4: the distilled student and the image-label control receive exactly the same image-label signal, so the only supervision difference between them is the presence of the Hungarian-matched box targets.

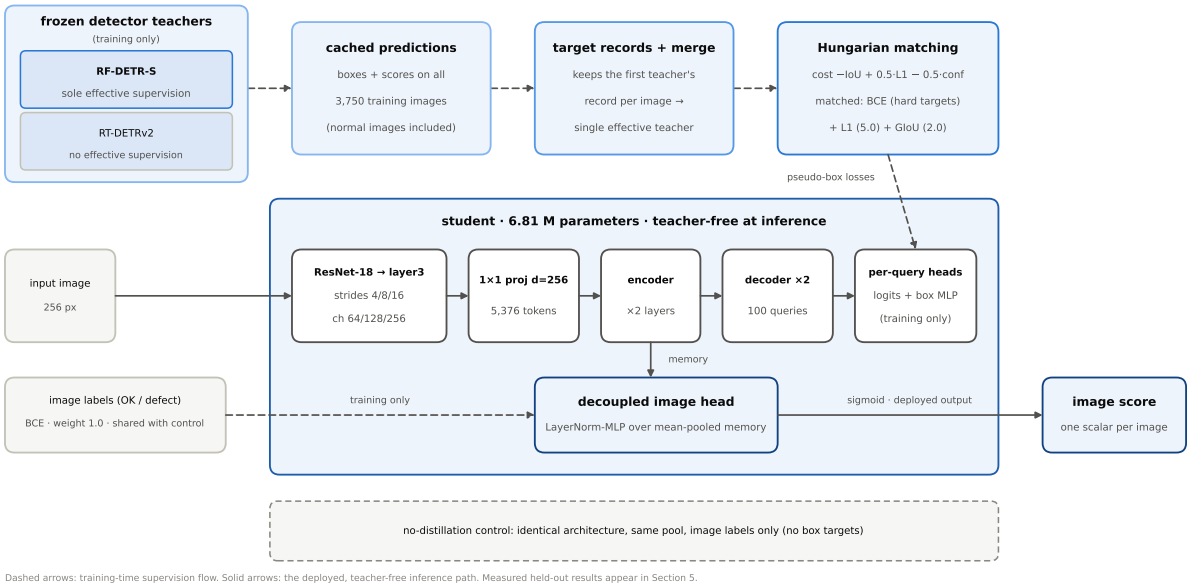

Figure 1: The box→score distillation pipeline as implemented. Frozen detector teachers are queried once; their cached predictions become per-image pseudo-box target records; the merge keeps the first teacher's record per image, so the strongest teacher is the sole effective supervision source in the reported runs; Hungarian matching assigns the targets to the student's queries, whose losses train the multi-scale student jointly with the image-label BCE consumed by the decoupled image head. At inference the student runs teacher-free and emits only the image head's score. The measured results appear in Section 5.

4. Experimental protocol

4.1 Dataset and evaluation surface

All experiments use MVTec AD (Bergmann et al. 2019) under a pooled-category protocol: a single model is trained over all fifteen texture and object categories and scored once, on one held-out evaluation surface of 802 images (614 normal / 188 defective). This is not the per-category protocol of the published MVTec AD literature, so our numbers are not comparable to published per-category leaderboards; the in-protocol

reference points are the rows of the landscape table in Section 5, all measured on this same surface. The evaluation surface and its disjointness machinery pre-date this study; this paper consumes them unchanged and contributes nothing to their construction.

4.2 Training pool and disjointness

Both the distilled student and the image-label control train on the box-annotated pool of 3,750 images (2,868 normal / 882 defective). The teacher checkpoints that generated the distillation targets pre-date this study: they were trained not on this pool but on a broader box-annotated detection split of the same dataset, and that split is not disjoint from the evaluation surface — 711 of the 802 evaluation images appear in its training portion. The student never sees that split: its supervision consists solely of the teachers' cached predictions over the 3,750-image pool, consumed for 30,000 steps (Section 3), plus the pool's image labels. The student's own train/test disjointness therefore holds, and is enforced at the image-identifier level — file-name checks are not sufficient — with the train∩test intersection re-verified to be exactly zero before the control run was launched. The teachers' evaluation-surface overlap nonetheless remains an indirect channel from that surface into the distillation targets, disclosed as part of the teacher-dependence limitation (Section 9). The teacher rows of Section 5 are measured from separate, later checkpoints trained on the pool itself and free of this overlap; they are not the checkpoints that produced the distillation targets.

4.3 The image-label control

The control is the paper's central design choice: the identical student architecture trained directly on the same pool's binary image labels, with the box annotations withheld — the natural "just train a classifier on the same data" reading of the recipe, and the strongest control available under the architecture's previously validated image-label recipe. That recipe is used unmodified (1,500 steps, batch size 16, AdamW with learning rate 1e-4 and weight decay 1e-4); its image-level binary cross-entropy is the only label-dependent loss — with no box targets, the recipe's remaining query-classification term reduces to a label-independent background regulariser on the query logits. The schedule asymmetry against the distilled student's 30,000 steps is disclosed here and treated in Sections 7 and 9, where a matched-schedule control is recorded as untested. The decision rule was pre-registered before the control ran, over three planned control seeds: distillation would be claimed useful only if it beat every control seed by more than 0.01 AUROC. Two of the three control seeds ran (1337 and 2027); the third was deferred (Section 4.4), so the rule is evaluated over the two control seeds that ran.

4.4 Metrics and statistics

The headline metric is threshold-free image-level AUROC, reported alongside AUPR. False-alarm and missed-defect rates at a fixed threshold of 0.5 are reported descriptively only; the threshold is never tuned. The distilled student is a single-seed measurement (seed 1337, the seed shared with the first control run); the control is measured at two seeds (1337 and 2027), with the third planned seed deferred. The central comparison is quantified by a paired bootstrap over the shared evaluation images (1,000 resamples, percentile 95% confidence intervals) between each control seed's saved predictions and the distilled student's; training-seed variability on the student's side remains unmeasured (Section 9). All metrics are recomputed from saved prediction files, never from transient training state.

5. Main results

This section reports the paper's numbers in three steps: the pre-registered control comparison, the full supervision landscape on the shared surface, and the retention–compression trade-off. All values are computed from saved predictions over the single fixed held-out evaluation surface of Section 4.

5.1 The falsification test

The central comparison is the control defined in Section 4: the identical student architecture trained on the same box-annotated pool with the boxes withheld, receiving exactly the image-label signal that the distilled student also receives. The control reaches 0.9114 AUROC and 0.8311 AUPR on the held-out surface at seed 1337, and 0.9320 and 0.8612 at seed 2027; the distilled scout reaches 0.9614 and 0.9323. The AUROC margins are +0.050 over the shared-seed control and +0.029 over the second — five and about three times the pre-registered 0.01 indistinguishability bar — and the AUPR margins are larger still, +0.101 and +0.071. A paired bootstrap over the 802 shared evaluation images (1,000 resamples; Section 4) places the 95% confidence interval of the AUROC margin at [0.030, 0.071] against the seed-1337 control and [0.010, 0.048] against the seed-2027 control, with the distilled scout ahead in 100% and 99.8% of resamples respectively. The distilled student itself remains one training run; Section 7 discusses what this evidence does and does not support.

At the frozen decision threshold of 0.5 — reported descriptively, never tuned (Section 4) — the distilled scout produces 342.0 false alarms per 10,000 normal images at a missed-defect rate of 0.1596, against 472.3 and 0.3777 for the seed-1337 control and 260.6 and 0.3670 for the seed-2027 control.

The control's absolute level is itself informative: its two seeds, 0.9114 and 0.9320, land inside the band spanned by the weak-label scout rows of Table 1 (0.9110 balanced, 0.9339 few-defect) — the same architecture trained on far smaller image-label budgets. Feeding this architecture the full pool's image labels buys essentially nothing over those much cheaper training sets — the image-label signal saturates. More image-label data alone does not close the gap to the distilled student; the box-level teacher targets are the variable that moves the needle.

5.2 The supervision landscape

Table 1 places both numbers in the supervision landscape measured on the same held-out surface (802 images: 614 normal, 188 defective). The ordering is stable and easy to read: the box-supervised detectors sit on top — RF-DETR-S at 0.9977 and D-FINE-S (Peng et al. 2024) at 0.9745 — the distilled scout (0.9614) sits between its two teachers, above RT-DETRv2 (0.9344), and every alternative that consumes cheaper supervision sits below it: the weak-label scout at 0.9339 (few-defect) and 0.9110 (balanced), the OK-only floor PatchCore (Roth et al. 2022) at 0.9242, and the image-label control at 0.9114. The AUPR column preserves the same ordering around the student, with a wider spread in the lower half of the table.

Method	Supervision consumed	Role	AUROC	AUPR
RF-DETR-S	boxes	teacher (strongest)	0.9977	0.9961
D-FINE-S	boxes	box-supervised reference	0.9745	0.9655
Distilled scout (ours)	teacher pseudo-boxes + image labels	student	**0.9614**	**0.9323**
RT-DETRv2	boxes	teacher (second)	0.9344	0.8966
Weak-label scout (few-defect)	image labels (smaller budget)	comparison	0.9339	0.8715
PatchCore	normal images only	OK-only floor	0.9242	0.8507
Image-label control (ours)	image labels (full pool)	control	**0.9114**	**0.8311**
Weak-label scout (balanced)	image labels (smaller budget)	comparison	0.9110	0.8019

Table 1 — Image-level AUROC and AUPR on the shared held-out evaluation surface. Every row is a single-seed measurement scored on the same 802 images; the two rows marked "ours" form the paper's controlled comparison pair, and the control row is its seed-1337 run — the control's second measured seed (0.9320) is reported in Section 5.1. The supervision column names the signal each method consumed at training time.

We concede the top of the table plainly: the distilled student does not beat its strongest teacher, and nothing in this paper claims it does. The claim is the ordering underneath — on this surface, no method that trains on less than box-level supervision reaches the distilled scout. The scout itself never sees a ground-truth box: it trains on teacher predictions and image labels, and the box-annotation cost sits upstream in its teachers, a scoped dependence stated in Section 2 and revisited in Section 9.

5.3 Retention and compression

The distilled scout retains 96.4% of its strongest teacher's AUROC (0.9614 vs 0.9977). In parameters — counted as the sum of checkpoint tensor elements, buffers included (Section 3) — the student is 6.81 M against 31.8 M for RF-DETR-S, roughly 4.7× fewer, and 20.5 M for RT-DETRv2, about 3.0× fewer, while scoring above the latter. Table 2 summarises the trade-off.

Model	Parameters	AUROC	Retention vs. strongest teacher
RF-DETR-S (teacher)	31.8 M	0.9977	100% (reference)
RT-DETRv2 (teacher)	20.5 M	0.9344	93.7%
Distilled scout (ours)	**6.81 M**	**0.9614**	**96.4%**

Table 2 — Parameters versus AUROC for the student and its two teachers. Parameter counts are sums of checkpoint tensor elements, buffers included (Section 3); retention is each model's AUROC as a fraction of the strongest teacher's 0.9977.

Parameter compression is the only efficiency claim this paper makes. The student's wall-clock latency was not profiled, no throughput measurement exists, and no speed advantage is claimed or implied; Section 9 records this as an open gap and Section 10 lists the deployment-oriented re-architecture as future work.

6. Historical cross-dataset consistency

The main result of Section 5 lives on a single benchmark, and Section 9 records that as a limitation. This section adds the only other evidence the project holds: four additional benchmark surfaces on which the same recipe was instantiated per dataset — detector teachers trained on that dataset's box-annotated pool, their cached predictions distilled into a fresh student — before a project-wide dataset freeze restricted all subsequent experiments to the benchmark of Section 4. These cells are *secondary* evidence: historical, single-seed, and closed to further runs; we draw no per-dataset conclusions from them.

6.1 Pre-freeze measurements (historical; no new runs; consistency evidence only)

Table 3 lists the five cells. Each is scored on a held-out test surface verified disjoint from that dataset's training pool at the image level, and each carries a clean machine-checked train/eval-disjointness verdict. The four non-anchor cells — MPDD (Jezek et al. 2021) at 0.9921, MVTec-AD-2 (Heckler-Kram et al. 2025) at 0.9703, VisA (Zou et al. 2022) at 0.8772, and Real-IAD (C. Wang et al. 2024) at 0.9426 — were measured before the freeze and are reported exactly as measured then. The VisA value carries a known score-polarity caveat on that surface and is reported verbatim, with no post-hoc correction.

One correction in this history is worth stating in full, because it illustrates why the disjointness verdicts matter. An earlier evaluation surface for three of these datasets (MPDD, MVTec-AD-2, VisA) overlapped the distillation training pool; re-scoring the same checkpoints on disjoint surfaces moved VisA from 0.9464 to 0.8772, while MPDD and MVTec-AD-2 moved only marginally. Only the disjoint-surface values appear in this paper — the inflated VisA figure is quoted here solely as the measured size of that inflation.

Benchmark surface	Image-AUROC	Status
MVTec-AD (anchor)	0.9614	current main result (Section 5)
MPDD	0.9921	historical, pre-freeze
MVTec-AD-2	0.9703	historical, pre-freeze
VisA	0.8772	historical, pre-freeze; score-polarity caveat
Real-IAD	0.9426	historical, pre-freeze

Table 3 — Image-AUROC of the distilled student across five benchmark surfaces, one recipe instantiation per dataset, each scored on a held-out surface disjoint from that dataset's training pool (clean disjointness

verdict; single seed). The four non-anchor rows are historical: measured before the project-wide dataset freeze, with no new runs for this paper. The VisA row carries a score-polarity caveat and is reported verbatim. No comparisons to other methods are made on these surfaces.

Read for what it is, the evidence is consistent rather than conclusive: across five industrial inspection benchmarks the recipe scores in a 0.877–0.992 band — the behaviour one expects of a mechanism rather than of an artefact of one dataset — while remaining what it is: historical, single-seed, secondary evidence (Section 9).

7. Analysis: what the gain is — and what it is not

What the gain is. An attribution result: the control of Section 4 shares the student's architecture, training pool, optimizer, input resolution, and image-label signal, so the only supervision difference is the Hungarian-matched teacher-box targets. The Section 5 separation — repeated on both measured control seeds — is evidence that box-level detector knowledge transfers into image-level anomaly scoring — under the disclosed protocol: one training run for the student, two control seeds, one benchmark for the main result.

What it is not. Not localisation: the student's boxes, a training by-product, do not localise (box Recall@Budget AUC 0.104 in the project's audit); only the image score is deployed. Not speed: the student is unprofiled; efficiency claims are parameters only (Section 5). Not detector-beating: 0.9614 trails the strongest teacher's 0.9977, conceded in Section 5. Not full multi-seed statistics: the distilled student is one training run and the third planned control seed is deferred; the paired bootstrap of Section 5 keeps both margins positive but captures evaluation-resampling uncertainty on the fixed surface, not training-seed variability of the student itself.

Corroborations (supporting, not upgrading). Same-architecture, same-surface contrasts: ground-truth boxes give 0.5584 (multi-scale) and 0.6128 (single-scale) versus 0.9614 with teacher boxes, which, unlike ground truth, cover all training images, normals included — a collapse not fully explained, possibly training dynamics. A historical paired with/without-teacher ablation on an audit-grade surface overlapping the training pool gave direction-consistent per-seed gains of +0.0255–0.0643 (sign-agreement $\geq$99.9%). A cross-dataset teacher swap collapsed the student by 0.38–0.49 AUROC on pre-cleanup surfaces — teacher provenance, not machinery alone, is load-bearing.

Reductions, pre-empted. "Pseudo-labeling, not distillation": conceded (hard targets, Section 3); the contribution is the cross-task formulation plus controlled attribution, not a new loss. "The control trained fewer steps": disclosed in Section 4; its unmodified recipe lands at the weak-label scout level on both seeds — signal saturation, not under-training (matched-schedule control: future work). "OK-only methods score higher": different protocol and supervision (in-protocol floor: Section 5).

8. Related work

Knowledge distillation and detector KD. Distillation began as soft-target transfer from large models into small ones (Hinton, Vinyals, and Dean 2015). In the DETR family, Hungarian-matched teacher targets are standard machinery — used by DETRDistill (Chang et al. 2023) and generalised by KD-DETR (Y. Wang et al. 2024) — but always detector-to-detector. We claim no novelty for the matching mechanism (Section 3); our reuse crosses task and granularity: boxes supervise an image-level score.

Cross-task distillation. The closest precedent, MLIC-KD-WSD (Liu et al. 2018), distils a weakly-supervised detection teacher into a multi-label image classifier — the same region-to-image direction, but general-domain, with a teacher itself trained without boxes. The reverse crossing (Guo, Alvarez, and Salzmann 2021) and the classification/detection protocol mismatch (Yang et al. 2023) show the task boundary is actively studied; score distillation for anomaly detection (Hong and Kang 2024) shares our teacher-free deployment story, with a task-preserving, unsupervised teacher. To our knowledge, distilling a box-supervised defect detector into a teacher-free image-level anomaly scorer via Hungarian-matched targets is unpublished.

OK-only distillation in industrial anomaly detection. STPM (G. Wang et al. 2021), Reverse Distillation (Deng and Li 2022), EfficientAD (Batzner, Heckler, and König 2024), and DeSTSeg (Zhang et al. 2023) fit a student to a frozen, task-agnostic feature network on normal images only; the teacher–student discrepancy is the anomaly signal. Supervision and mechanism are opposite to ours — no defect labels, teacher as reference, residual as score. Our teacher is task-supervised on the same dataset's defects and the deployed score is a learned discriminative head.

Industrial benchmarks and supervised defect detection. The MVTec-AD family (Bergmann et al. 2019; Heckler-Kram et al. 2025), VisA (Zou et al. 2022), MPDD (Jezek et al. 2021), and Real-IAD (C. Wang et al. 2024) anchor industrial evaluation; PatchCore (Roth et al. 2022) is the OK-only reference, KolektorSDD2 the mixed weak-to-full supervision precedent (Božič, Tabernik, and Skočaj 2021), and deployment-reality critiques motivate compact front-line deployables (Baitieva et al. 2025). Our teachers come from the real-time DETR line (Carion et al. 2020; Zhao et al. 2024; Lv et al. 2024; Peng et al. 2024; Robinson et al. 2026) — detectors this paper treats as teachers to compress across tasks, not as rivals to outscore.

9. Limitations

The central comparison is seed-limited. The distilled student is a single training run (seed 1337); the control ran at two of its three planned seeds (1337 and 2027), with the third deferred. The paired bootstrap of Section 5 quantifies scoring variability over the fixed evaluation surface, not training-seed variability of the distilled student, and the margin over the stronger control seed (+0.029) is the conservative reading of the central result.

The main result is single-benchmark. Every headline number lives on one dataset and one held-out surface (Section 4), measured under a project-wide dataset freeze; the cross-dataset cells of Section 6 are historical measurements, not new evidence, and no claim extends beyond the surface it was measured on.

Latency is unprofiled. The student's wall-clock speed was never measured, on any hardware, at any batch size; parameter count is the paper's only efficiency statement, and nothing here should be read as a speed claim. A deployment-oriented faster-student re-architecture is future work.

The dual-teacher merge is effectively single-teacher. As implemented, the merge keeps the first teacher's record per image, and because that teacher's predictions cover every training image, the second teacher contributed no supervision in the reported runs (Section 3). Per-teacher and true-union ablations are future work; the reported result should be read as distillation from one detector.

Targets are hard. Teacher confidences are not propagated into the loss; soft-target distillation is untested.

The recipe is teacher-dependent. It presupposes a box-annotated pool and a detector trained on it: annotation cost is moved upstream, not eliminated. Historical teacher-provenance evidence (Section 7) indicates the teacher's training distribution is load-bearing.

No localisation capability is claimed. The student's boxes are a training by-product and do not localise (Section 7); the deployed output is the image-level score alone.

The control's recipe was not tuned for this comparison. The control used its unmodified validated recipe (1,500 steps), far shorter than the distilled schedule (Section 4). Section 7 argues the image-label signal saturates rather than under-trains, but a matched-schedule control — and, more broadly, recipe tuning that might narrow the gap — is untested.

The protocol is pooled-category. One model is trained over all categories; numbers are not comparable to the published per-category results on this benchmark (Section 4).

The ground-truth-box contrast is unexplained. The same architecture collapses when trained on ground-truth boxes instead of teacher predictions (Section 7); the collapse is observed, not explained.

10. Conclusion

Box-supervised detector knowledge was distilled across tasks into a compact, teacher-free image-level anomaly scout, and the controlled same-architecture comparison attributes the gain to the box targets: +0.050 and +0.029 AUROC over the two measured control seeds, both beyond the pre-registered bar (Section 5) — more image-label data alone did not close the gap. The student retained 96.4% of its strongest teacher's AUROC at roughly 4.7× fewer parameters, and historical pre-freeze measurements on four further surfaces fall in a 0.877–0.992 band (Section 6). The claim stops there: no localisation, no wall-clock speedup, no advantage over the teacher, and no full multi-seed statistics — the distilled student is one training run on one benchmark, with one teacher the sole effective supervision source (Section 9). The deferred work follows directly: seed replication on both sides, per-teacher and soft-target ablations, a matched-schedule control, latency profiling with a faster deployment-oriented student, and re-opened multi-dataset evaluation. Within these bounds the answer is affirmative: boxes taught scores, and the transfer ran across tasks.

References

Baitieva, Aimira, Yacine Bouaouni, Alexandre Briot, Dick Ameln, Souhaiel Khalfaoui, and Samet Akcay. 2025. "Beyond Academic Benchmarks: Critical Analysis and Best Practices for Visual Industrial Anomaly Detection." https://arxiv.org/abs/2503.23451.

Batzner, Kilian, Lars Heckler, and Rebecca König. 2024. "EfficientAD: Accurate Visual Anomaly Detection at Millisecond-Level Latencies." In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. https://arxiv.org/abs/2303.14535.

Bergmann, Paul, Michael Fauser, David Sattlegger, and Carsten Steger. 2019. "MVTec AD – a Comprehensive Real-World Dataset for Unsupervised Anomaly Detection." In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://www.mvtec.com/research-teaching/datasets/mvtec-ad.

Božič, Jakob, Domen Tabernik, and Danijel Skočaj. 2021. "Mixed Supervision for Surface-Defect Detection: From Weakly to Fully Supervised Learning." *Computers in Industry*. https://arxiv.org/abs/2104.06064.

Carion, Nicolas, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. "End-to-End Object Detection with Transformers." In *European Conference on Computer Vision (ECCV)*. https://arxiv.org/abs/2005.12872.

Chang, Jiahao, Shuo Wang, Haiming Xu, Zehui Chen, Chenhongyi Yang, and Feng Zhao. 2023. "DETRDistill: A Universal Knowledge Distillation Framework for DETR-Families." In *IEEE/CVF International Conference on Computer Vision (ICCV)*. https://arxiv.org/abs/2211.10156.

Cheng, Yuqi, Yunkang Cao, Haiming Yao, Wei Luo, Cheng Jiang, Hui Zhang, and Weiming Shen. 2025. "A Comprehensive Survey for Real-World Industrial Defect Detection: Challenges, Approaches, and Prospects." https://arxiv.org/abs/2507.13378.

Deng, Hanqiu, and Xingyu Li. 2022. "Anomaly Detection via Reverse Distillation from One-Class Embedding." In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://arxiv.org/abs/2201.10703.

Guo, Shuxuan, Jose M. Alvarez, and Mathieu Salzmann. 2021. "Distilling Image Classifiers in Object Detectors." In *Advances in Neural Information Processing Systems (NeurIPS)*. https://arxiv.org/abs/2106.05209.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. "Deep Residual Learning for Image Recognition." In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Heckler-Kram, Lars, Jan-Hendrik Neudeck, Ulla Scheler, Rebecca König, and Carsten Steger. 2025. "The MVTec AD 2 Dataset: Advanced Scenarios for Unsupervised Anomaly Detection." https://arxiv.org/abs/2503.21622.

Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. 2015. "Distilling the Knowledge in a Neural Network." https://arxiv.org/abs/1503.02531.

Hong, Jeongmin, and Seokho Kang. 2024. "Score Distillation for Anomaly Detection." *Knowledge-Based Systems* 295. https://doi.org/10.1016/j.knosys.2024.111842.

Jezek, Stepan, Martin Jonak, Radim Burget, Pavel Dvorak, and Milos Skotak. 2021. "Deep Learning-Based Defect Detection of Metal Parts: Evaluating Current Methods in Complex Conditions." In *International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*. https:

//github.com/stepanje/MPDD.

Liu, Yongcheng, Lu Sheng, Jing Shao, Junjie Yan, Shiming Xiang, and Chunhong Pan. 2018. "Multi-Label Image Classification via Knowledge Distillation from Weakly-Supervised Detection." In *ACM International Conference on Multimedia (ACM MM)*. https://arxiv.org/abs/1809.05884.

Lv, Wenyu, Yian Zhao, Qinyao Chang, Kui Huang, Guanzhong Wang, and Yi Liu. 2024. "RT-DETRv2: Improved Baseline with Bag-of-Freebies for Real-Time Detection Transformer." https://arxiv.org/abs/2407.17140.

Oquab, Maxime, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, et al. 2023. "DINOv2: Learning Robust Visual Features Without Supervision." https://arxiv.org/abs/2304.07193.

Peng, Yansong, Hebei Li, Peixi Wu, Yueyi Zhang, Xiaoyan Sun, and Feng Wu. 2024. "D-FINE: Redefine Regression Task in DETRs as Fine-Grained Distribution Refinement." https://arxiv.org/abs/2410.13842.

Robinson, Isaac, Peter Robicheaux, Matvei Popov, Deva Ramanan, and Neehar Peri. 2026. "RF-DETR: Neural Architecture Search for Real-Time Detection Transformers." In *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/2511.09554.

Roth, Karsten, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. 2022. "Towards Total Recall in Industrial Anomaly Detection." In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://arxiv.org/abs/2106.08265.

Wang, Chengjie, Wenbing Zhu, Bin-Bin Gao, Zhenye Gan, Jiangning Zhang, Zhihao Gu, Shuguang Qian, Mingang Chen, and Lizhuang Ma. 2024. "Real-IAD: A Real-World Multi-View Dataset for Benchmarking Versatile Industrial Anomaly Detection." In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://arxiv.org/abs/2403.12580.

Wang, Guodong, Shumin Han, Errui Ding, and Di Huang. 2021. "Student-Teacher Feature Pyramid Matching for Anomaly Detection." In *British Machine Vision Conference (BMVC)*. https://arxiv.org/abs/2103.04257.

Wang, Yu, Xin Li, Shengzhao Weng, Gang Zhang, Haixiao Yue, Haocheng Feng, Junyu Han, and Errui Ding. 2024. "KD-DETR: Knowledge Distillation for Detection Transformer with Consistent Distillation Points Sampling." In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16016–25. https://arxiv.org/abs/2211.08071.

Yang, Longrong, Xianpan Zhou, Xuewei Li, Liang Qiao, Zheyang Li, Ziwei Yang, Gaoang Wang, and Xi Li. 2023. "Bridging Cross-Task Protocol Inconsistency for Distillation in Dense Object Detection." In *IEEE/CVF International Conference on Computer Vision (ICCV)*. https://arxiv.org/abs/2308.14286.

You, Zhiyuan, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. 2022. "A Unified Model for Multi-Class Anomaly Detection." In *Advances in Neural Information Processing Systems (NeurIPS)*. https://arxiv.org/abs/2206.03687.

Zhang, Xuan, Shiyu Li, Xi Li, Ping Huang, Jiulong Shan, and Ting Chen. 2023. "DeSTSeg: Segmentation Guided Denoising Student-Teacher for Anomaly Detection." In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://arxiv.org/abs/2211.11317.

Zhao, Yian, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. 2024. "DETRs Beat YOLOs on Real-Time Object Detection." In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://arxiv.org/abs/2304.08069.

Zou, Yang, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. 2022. "SPot-the-Difference Self-Supervised Pre-Training for Anomaly Detection and Segmentation." In *European Conference on Computer Vision (ECCV)*. https://github.com/amazon-science/spot-diff.