

Systematic Diagnosis of Cross-Domain Performance Degradation in PCB Defect Detection: A PKU-Source Study

Huanle Yu

Guizhou University of Engineering Science

Dechrit Maneetham

`dechrit_m@rmutt.ac.th`

Rajamangala University of Technology Thanyaburi

Shenghan Gao

Rajamangala University of Technology Thanyaburi

Xia Luo

Bijie Municipal Rural Revitalization Development Center

Weihua Zhang

Guizhou University of Engineering Science

Yulan Gao

Guizhou University of Engineering Science

Xuefeng Ding

Guizhou University of Engineering Science

Research Article

Keywords: PCB defect detection, domain shift, industrial machine vision, annotation consistency, imaging sensitivity, cross-domain evaluation

Posted Date: September 29th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10939592/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Systematic Diagnosis of Cross-Domain Performance Degradation in PCB Defect Detection: A PKU-Source Study

Huanle Yu¹, Dechrit Maneetham^{2*}, Shenghan Gao², Xia Luo³,
Weihua Zhang¹, Yulan Gao¹, Xuefeng Ding¹

¹ College of Mechanical Engineering, Guizhou University of Engineering Science, Bijie, Guizhou 551700, China

² Department of Mechatronics Engineering, Rajamangala University of Technology Thanyaburi, Thanyaburi, Pathum Thani 12110, Thailand

³ Bijie Municipal Rural Revitalization Development Center, Bijie, Guizhou 551700, China

* Corresponding author.

E-mail: Huanle Yu: huanleyu@gues.edu.cn; Dechrit Maneetham: dechrit_m@rmutt.ac.th; Shenghan Gao: shenghan_g@rmutt.ac.th; Xia Luo: luoxia9935@163.com; Weihua Zhang: zhangweihua@gues.edu.cn; Yulan Gao: yulangao@gues.edu.cn; Xuefeng Ding: xuefengding@gues.edu.cn

ORCID: Huanle Yu: 0009-0007-4885-4450; Dechrit Maneetham: 0000-0002-5635-4153

Author contributions

Huanle Yu: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Visualization, Writing—original draft.

Dechrit Maneetham: Conceptualization, Methodology, Supervision, Writing—review and editing.

Shenghan Gao: Validation, Formal analysis, Writing—review and editing.

Xia Luo: Investigation, Resources, Writing—review and editing.

Weihua Zhang: Investigation, Data curation, Validation, Writing—review and editing.

Yulan Gao: Investigation, Validation, Writing—review and editing.

Xuefeng Ding: Formal analysis, Visualization, Validation, Writing—review and editing.

All authors read and approved the final manuscript.

Abstract

Printed circuit board (PCB) defect detectors can achieve strong performance within their development domain yet degrade severely when transferred to heterogeneous inspection domains. This study systematically diagnoses such cross-domain degradation using PKU as the sole source domain and MIXED-clean, SELF, and DeepPCB as observational target domains under a common five-class transfer ontology. Two primary source-trained detectors were analyzed without target-domain adaptation, while RT-DETR-L provided full-domain cross-architecture confirmation. After PCB-template-disjoint source retraining, source-domain mean average precision (mAP₅₀₋₉₅) ranged from 0.4466 to 0.4827, whereas the maximum mean target-domain retention was 2.61% at an input size of 960 pixels. Within the locked 200-image audit subset, independent five-class validation and adjudicated Recall rescoring did not alter the severe source-to-target Recall pattern. Controlled imaging perturbations revealed local sensitivity where reference-condition detectability remained measurable, while imaging attribution remained unresolved for floor-limited target domains. The target domains also differed from the source in image appearance, PCB background, defect geometry, scale distribution, and dataset composition. For the primary models, missed-target states dominated the selected-target analysis across nine prespecified inference settings. Taken together, the observed PKU-to-target degradation co-occurred with multiple measurable domain mismatches, while no examined diagnostic branch alone provided a sufficient account. This study provides a traceable diagnosis of these source-to-target failures and a basis for broader validation.

Keywords: PCB defect detection; domain shift; industrial machine vision; annotation consistency; imaging sensitivity; cross-domain evaluation

1. Introduction

Printed circuit boards (PCBs) are fundamental components of modern electronic systems, and reliable visual inspection is essential for preventing manufacturing defects from propagating into downstream assembly and

product operation [1, 2]. Deep-learning-based object detectors have substantially improved the automation of PCB inspection [3], with recent studies focusing on lightweight architectures [4], multi-scale feature fusion, attention mechanisms, complex-background detection [5], and few-shot learning [6]. These developments have achieved strong detection performance on established PCB benchmarks and have improved the recognition of small or difficult defects under the conditions represented by their training and evaluation data.

However, strong performance within a benchmark or development environment does not necessarily imply reliable deployment under a different inspection domain. Much of the recent PCB-detection literature continues to emphasize architectural improvement and in-domain accuracy [4, 5, 7], often evaluating models on splits derived from the same dataset or under closely related acquisition conditions. Taken together, these studies demonstrate continuing progress in detector design, but provide much less evidence about what happens when a source-trained detector encounters a different PCB population [8], background appearance [9], defect-scale distribution [8], or acquisition condition [10]. This distinction is important for industrial deployment because a detector may retain the same class definitions while the visual and compositional characteristics of the target data differ substantially from those encountered during model development.

The broader object-detection literature recognizes this problem as domain shift. Recent domain-adaptation [11, 12] and single-source domain-generalization studies [13] have shown that changes between source and target data distributions can substantially impair detection performance, motivating methods based on feature alignment, data diversification, target-domain information, and domain-invariant representations. Industrial inspection is also sensitive to image-formation conditions [14, 15]; for example, multi-illumination quality-control research has demonstrated that illumination itself can materially affect defect segmentation and inspection behavior [16]. Yet adaptation and generalization methods mainly address how to improve transfer, whereas a more basic practical question remains insufficiently addressed in PCB inspection: what characterizes a source-trained detector’s failure after transfer, and which observable properties are associated with that failure?

Answering this question requires more than reporting a source-to-target accuracy drop. A low target-domain score can coexist with annotation disagreement [17, 18], sensitivity to image formation [14, 15], different defect-scale mixtures [8], changes in PCB background or object geometry [9, 10], and different detection-level failure mechanisms. These possibilities should not be treated as interchangeable. A detector that fails because the target is never detected presents a different problem from one that detects the target but localizes it poorly or assigns the wrong class. Likewise, stability under a small set of synthetic brightness or blur perturbations does not establish robustness to a real target domain whose background [9, 19], defect composition, and geometry may differ simultaneously. A useful diagnosis therefore needs to separate the existence of cross-domain degradation from the evidence used to investigate its possible explanations.

This study addresses that need through a systematic diagnostic workflow for the examined PKU-source transfers. PKU is used as the sole source domain, while MIXED-clean, SELF, and DeepPCB provide heterogeneous observational targets under a verified common five-class transfer ontology. Source-trained detectors are frozen before target evaluation so that target-domain data do not influence model selection or target-specific inference-parameter tuning. Full-domain evaluation, including RT-DETR-L cross-architecture confirmation, first establishes the transfer gap. The diagnostic analysis then examines annotation reliability and Recall impact, paired responses to controlled brightness, contrast, and Gaussian blur, defect-scale characteristics, and selected-target failure transitions. The diagnosis is further extended beyond selected targets through full-domain characterization of image appearance, PCB background, object geometry, density, scale composition, and class composition, together with multivariate domain analysis. Finally, domain–class–size stratification is used to characterize missed-target, localization-failure, and classification-confusion states among selected targets. Together, these components define the diagnostic workflow detailed in Sections 3 and 4.

The methodological contribution is a traceable source-to-target diagnostic protocol rather than a new detector architecture. It separates full-domain transfer measurement from fixed-instance diagnosis, checks annotation consistency and evaluation-chain integrity, and integrates controlled imaging sensitivity, post hoc domain characterization, and selected-target failure-state analysis under explicit metric and causal boundaries. Importantly, the heterogeneous datasets are treated as observed dataset domains rather than as a controlled factorial experiment. The resulting evidence is therefore used to identify consistent patterns and explanatory boundaries, not to assign additive causal percentages across the diagnostic evidence streams. Accordingly, the empirical claims are restricted to the examined PKU-source transfers.

2. Related Work

2.1 PCB Defect Detection and Industrial Visual Inspection

PCB visual inspection has progressed from rule-based image processing and handcrafted feature engineering toward data-driven deep learning. Earlier approaches relied heavily on image subtraction, template matching, segmentation, manually designed descriptors, and conventional classifiers [1, 2], whereas recent systems increasingly use convolutional and Transformer-based architectures to learn defect representations directly from inspection images [3]. Existing reviews show that this transition has substantially broadened the capability of automated PCB inspection, particularly for localization and recognition of complex defects that are difficult to describe using fixed visual rules [1, 2].

Recent research has consequently focused on improving the representation of small and visually subtle PCB defects. Multi-scale feature aggregation, attention mechanisms, high-resolution feature preservation, and lightweight detection heads have been introduced to reduce information loss and strengthen defect-background discrimination. LDD-Net and other lightweight YOLO-based detectors emphasize cross-scale feature fusion and computational efficiency [4], while more recent architectures combine multi-scale interaction with adaptive attention to improve sensitivity to micron-scale defects under complex textures [5]. The development of Transformer-based PCB detectors further illustrates that current progress is not confined to a single detector family [20, 21], but follows a broader trend toward stronger feature representation across scales and backgrounds.

Alongside architecture design, detector selection increasingly involves an explicit trade-off between accuracy and deployment efficiency [4]. Comparative evaluation of the YOLOv12 family on PCB defects showed that model scale changes the balance between detection performance and computational cost, reinforcing the need to select detectors according to the requirements of the inspection system rather than accuracy alone [22]. However, such evaluations and most architecture-oriented PCB studies primarily establish performance under the datasets and acquisition conditions represented during model development. Improving feature fusion, attention, or model efficiency does not by itself establish that the resulting detector will remain reliable when the PCB style, background, defect distribution, or acquisition condition changes.

A growing line of work has therefore begun to consider data efficiency and generalization beyond the original training setting. Zero-shot PCB detection, for example, has been explored as a means of reducing dependence on real annotated samples and adapting inspection to previously unseen PCB types and styles [9]. Such work reflects an important shift from optimizing only in-domain detection toward deployment under changing PCB conditions [8, 9]. Nevertheless, these approaches remain primarily method-oriented: they seek to improve generalization [9] or reduce adaptation cost [10, 23]. They provide less direct evidence about what characterizes source-trained detector failure after deployment to a heterogeneous target domain and which observable domain differences are compatible with that failure. This question motivates the domain-shift and robustness literature reviewed next.

2.2 Domain Shift and Robustness in Machine Vision

Domain shift arises when the data distribution encountered during deployment differs from that used for model development. In object detection, this mismatch can affect both feature representation and downstream localization or classification, motivating domain-adaptation [11, 12, 24] and domain-generalization methods [13] based on feature alignment, adversarial learning, contrastive representation learning, and data diversification. Similar concerns are particularly relevant to industrial inspection, where changes in imaging conditions, manufacturing environments, product appearance, background texture, and object scale can occur without changing the nominal inspection task. Existing industrial studies have therefore explored adaptation across production conditions [25], manufacturers [26], and unseen datasets rather than assuming that performance measured in the source domain will be preserved after deployment [13].

Recent PCB studies provide more direct evidence that such distribution changes are intrinsic to practical board inspection. Yan et al. formulated PCB continuous defect detection as a domain-incremental problem in which different production pipelines, product batches, manufacturing processes, equipment, and environmental conditions generate defect data with changing distributions [8]. Their experiments further treated HRIPCB, DeepPCB, and RoboflowPCB as distinct PCB domains whose image appearance, defect scale, acquisition characteristics, and composition differ substantially. Rather than assuming one stationary PCB distribution, this work explicitly separates domain-common and domain-specific knowledge so that the detector can adapt sequentially while retaining previously learned domains.

Other PCB studies have addressed related forms of domain mismatch through different adaptation strategies. Nguyen et al. showed that detectors trained on optical or synthesized PCB images can deteriorate when applied to real X-ray imagery because of modality differences, scanner configurations, background contexts, and changes in defect appearance; their feature-based adaptation framework was designed to reduce these discrepancies using limited target-domain data [10]. Wen et al. considered different PCB styles as distinct domains and used multi-

source feature alignment to improve target-domain defect classification, explicitly noting that working environment, machinery, and PCB style can alter the training–test distribution [23]. In contrast, Yin et al. pursued zero-shot generalization, arguing that differences among PCB types and styles make adaptation to unseen boards costly and using synthetic training with reference-guided detection to reduce dependence on target-domain annotation [9]. Together, these studies show that PCB domain shift can arise through several different mechanisms and that recent research has begun to move beyond purely in-domain detector optimization.

Controlled robustness studies provide a complementary perspective. Changes in brightness, noise, blur, color, or scale can reveal sensitivity to specific image degradations, and several industrial or PCB studies have incorporated such perturbations to examine model robustness. For example, the FC-DANN study [23] evaluated classification robustness under increasing Gaussian noise, while recent zero-shot PCB work explicitly included noise [9], motion blur, illumination, shadow, translation, and scale variation during training. These experiments are useful for isolating sensitivity to known perturbations, but they are not equivalent to evaluation on a heterogeneous real target domain, where acquisition appearance, background, defect geometry, scale, and composition may change simultaneously.

A common feature of these approaches is that they primarily address how to preserve or recover performance after domain shift has been recognized [27]. Domain adaptation aligns source and target representations [11, 12], incremental learning updates models as new PCB domains arrive [8], and zero- or few-shot methods reduce the amount of target-domain supervision required [6, 9]. What remains less developed is systematic diagnostic analysis before adaptation. Once a source-trained PCB detector has failed on heterogeneous target domains, it is still necessary to determine whether the observed degradation is associated with annotation inconsistency, local imaging sensitivity, scale and compositional differences, broader domain characteristics, or a particular detection-level failure pattern. The present study addresses this complementary problem by diagnosing the cross-domain performance gap without target-domain adaptation and without interpreting heterogeneous dataset differences as an additive causal decomposition.

2.3 Annotation Quality and Cross-Domain Evaluation

Reliable ground truth (GT) is a prerequisite for both detector training and cross-domain evaluation [17, 18, 28]. In object detection, annotation errors may involve missing objects, incorrect class assignments, inaccurate bounding-box positions or extents, and inconsistent treatment of ambiguous instances [17, 18]. Previous studies have shown that such errors can alter measured detector performance [17, 18] and, when present in training data, can also affect the representations learned by the model [18, 28]. Consequently, an observed performance difference between datasets cannot automatically be attributed to visual domain shift without considering whether the annotation conventions and annotation quality are sufficiently consistent for comparison.

This issue is particularly relevant when public datasets collected or labeled under different protocols are compared [17]. Dataset-quality studies have reported that correcting missing or inaccurate ground-truth annotations can materially change the evaluation of object-detection systems [17], while investigations of label noise have shown that different forms of annotation corruption do not affect detection models in the same way [18]. Industrial inspection introduces an additional difficulty because defect boundaries can be small, irregular, or visually ambiguous, making annotation quality dependent on both labeling rules and reviewer expertise [29]. Annotation consistency should therefore be treated as an independent source of evaluation uncertainty rather than being implicitly absorbed into the broader concept of domain shift.

Recent PCB evidence illustrates this distinction directly. Yin et al. identified annotation inconsistencies and omitted small defects in a public PCB dataset, including rectangular annotations that did not closely follow irregular defect morphology [9]. Their manual re-annotation changed the resulting evaluation, demonstrating that apparent detector error can partly reflect imperfections in the reference annotations rather than the prediction alone [9]. This observation is important for cross-dataset PCB studies because datasets can share nominal defect categories while differing in annotation detail, boundary conventions, or completeness. At the same time, evidence that annotation quality affects evaluation does not establish that annotation inconsistency is responsible for a particular cross-domain performance collapse.

Existing work has established that annotation quality matters [17], but cross-domain evaluation requires more than reviewer agreement: it must also establish whether the formal labels are sufficiently reliable and whether correcting them materially changes the measured transfer gap [18, 28]. The present study therefore combines inter-reviewer auditing with independent five-class ground-truth validation and frozen-prediction Recall rescoring.

3. Problem Formulation and Diagnostic Workflow

3.1 Problem Definition and Data Roles

This study investigates the behavior of PCB defect detectors when they are deployed outside the image domain used for model development. Let $\mathcal{D}_s$ denote the source domain and $\mathcal{D}_t^{(k)}$ the k -th target domain. PKU is used as the sole source domain, while MIXED-clean, SELF, and DeepPCB serve as heterogeneous target domains. Models are trained and selected using source-domain data only and are then frozen for evaluation on all four test domains. Target-domain annotations are not used to train or select these source-trained models.

Because the native label spaces differ, the formal transfer protocol uses only classes validated as exact mappings across all four domains (Table S1):

$$\mathcal{C} = \{\text{mouse_bite, open_circuit, short, spur, spurious_copper}\}.$$

DeepPCB native ID 5 was reviewed as pin-hole rather than PKU missing_hole and was excluded from Common-5 GT, yielding the formal $K = 5$ transfer task. All source-trained detector rows use this formal five-class transfer ontology. Tables S1-S8 are provided in Online Resource 1.

Here, cross-domain transfer denotes evaluation across heterogeneous PCB datasets that differ in acquisition appearance [10], PCB and defect composition [8, 9], and object scale [30]; annotation characteristics [18] are treated separately as evaluation uncertainty. Because the datasets were not collected as a controlled factorial study across physical production lines, they are treated as observed dataset domains, and no causal percentage of the performance gap is assigned to any individual factor.

Two complementary evaluation settings are used. Full-domain evaluation uses full-image Common-5 GT for the retained formal test population to establish the source-to-target performance gap; images with no Common-5 GT after ontology filtering are excluded (Table S2). A separate fixed-instance diagnostic branch evaluates a locked set of selected targets under paired controlled perturbations and stratifies the same observations by scale and error type. This branch characterizes detector behavior and does not replace full-domain evaluation. Together, the two settings separate the observed cross-domain performance gap from the diagnostic evidence used to investigate its possible explanations.

3.2 Diagnostic Workflow and Research Questions

The diagnostic workflow begins by establishing whether source-trained detectors consistently lose performance on heterogeneous target domains [8–10] and then examines the same degradation from four complementary perspectives: annotation reliability and Recall impact [17, 18], controlled imaging sensitivity [14, 15], domain and scale characteristics [8], and selected-target failure states. These evidence streams are interpreted jointly, but not as additive components of a causal decomposition (Fig. 1).

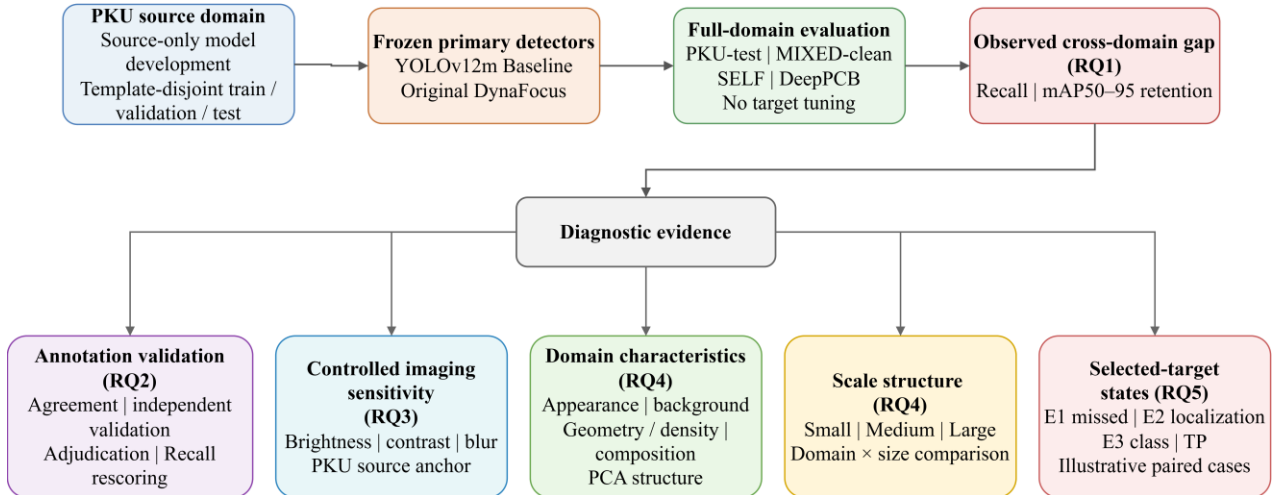


Fig. 1 Diagnostic workflow for the PKU-source cross-domain study; RQ4 analyses are post hoc

Five research questions structure the analysis:

RQ1: How severe and consistent is the performance degradation of PKU-trained PCB defect detectors across the three heterogeneous target domains?

RQ2: Following independent ground-truth validation, how much does adjudicated label correction change frozen-prediction Recall within the audited subset?

RQ3: How do fixed defect detections respond to controlled changes in brightness, contrast, and blur across the source anchor and target domains, and how does reference-condition detectability constrain sensitivity interpretation?

RQ4 (post hoc): How do defect-scale composition and measurable visual, geometric, and compositional characteristics differ between the source and target domains?

RQ5: Among selected targets, which GT-associated states—missed target, localization failure, or classification confusion—predominate under cross-domain transfer?

The questions form a progressive diagnostic sequence. RQ1 establishes the transfer problem using the full retained Common-5 test populations. RQ2 tests ground-truth reliability and quantifies its effect on frozen-prediction Recall within the audited subset. RQ3 tests paired responses to selected image-quality factors against a PKU source anchor, with sensitivity interpreted only where reference-condition detectability remains measurable. RQ4 provides post hoc characterization of broader domain differences without reference to detector predictions. RQ5 characterizes GT-associated failure states among selected targets; the evidence is integrated in Section 6. Sections 4.1–4.4 define the experimental and diagnostic procedures, Section 4.5 specifies the shared statistical analysis, and Sections 5.1–5.5 report the findings in the RQ1–RQ5 sequence.

4. Experimental Methodology

4.1 Datasets, Detectors, and Unified Evaluation Protocol

Four heterogeneous PCB datasets were assigned fixed experimental roles (Table 1). PKU served as the sole source domain for the primary source-trained models [31], whereas MIXED-clean [32], SELF, and DeepPCB [33] served as observational targets. An earlier candidate set also included DsPCBSD+, which was excluded from the final protocol because its native defect taxonomy was not sufficiently aligned with the shared defect definitions required for the Common-5 transfer ontology. The three retained domains constitute the observational target set; no population-level representativeness is assumed. SELF is a self-collected observational target acquired with an MV-CU060-10GC area-scan camera under controlled LED illumination; its public release is identified in the Data availability statement, and SELF was not used for source-domain model development or target-specific tuning. Target-domain transfer evaluation used the five-class ontology in Section 3.1; native mappings and filtering QC are reported in Tables S1–S2. Non-Common-5 boxes were removed without ignore regions; retained images therefore contained excluded visible defects as unlabeled distractors, whereas excluded-only images were omitted. Target-domain data were not used to train, select, or tune any source-trained detector. PKU was repartitioned by verified PCB-template group, assigning six/two/two templates and 481/92/120 original–rotation pairs to train/validation/test; Common-5 filtering retained 100 of the 120 test lineages (200 of 240 images) for formal evaluation. For this PKU source split, template overlap, pair-integrity failures, cross-split exact SHA-256 duplicates, and perceptual-hash (pHash) candidates at the locked distance threshold (≤ 6) were all zero; no authoritative physical-PCB identifier was available.

The primary comparison used YOLOv12m Baseline [22, 34] and the previously published Original DynaFocus detector [35]. Original DynaFocus is used only for architectural comparison; its development is not a contribution of this study. The configurations and reproducibility resources used here are available in the repository identified in the Code availability statement. For confirmation, COCO-pretrained RT-DETR-L [36, 37] used the same COCO-pretrained rtdetr-l.pt initialization for all three seeds and the same template-disjoint PKU Common-5 runtime configuration. Under Ultralytics 8.4.120, it was trained for 100 epochs at $\text{imgsz} = 960$ with $\text{batch}=2$ and no early stopping; $\text{optimizer}=\text{auto}$ resolved to AdamW ($\text{lr}=0.001111$, $\text{momentum}=0.9$, $\text{weight decay}=5 \times 10^{-4}$). Logged augmentation used $\text{HSV}=0.015/0.7/0.4$, $\text{translate}=0.10$, $\text{scale}=0.50$, $\text{fliplr}=0.50$, $\text{mosaic}=1.0$ with $\text{close_mosaic}=10$, $\text{MixUp}=0$, plus Blur, MedianBlur, ToGray, and CLAHE at $p=0.01$ each. Common-5 PKU-validation mAP_{50-95} selected the checkpoint for the four formal test domains. RT-DETR-L was used only for full-domain cross-architecture confirmation, not architecture ranking or fixed-instance analysis.

Common-5 checkpoints were frozen before target evaluation; their five-class outputs were evaluated without class filtering. All retained formal Common-5 test images were evaluated under the locked full-domain settings with full-image Common-5 GT to establish source-to-target performance degradation, with Precision, Recall, mAP_{50} , and mAP_{50-95} reported from that GT. The same frozen primary models were subsequently used for the fixed-instance imaging, scale, and error analyses described in Sections 4.3–4.4. Within the primary transfer and diagnostic analyses, no target-specific retraining, fine-tuning, sample reselection, or prediction-generation setting was chosen based on target performance; the evaluated models therefore remained source-only.

295 Table 1. Unified experimental and evaluation protocol

Item	Setting
Data roles	PKU: sole source domain for primary model development; MIXED-clean, SELF, and DeepPCB: observational targets for source-only evaluation, with non-test target data used only for isolated oracle controls.
Transfer ontology	5 classes: mouse_bite, open_circuit, short, spur, spurious_copper.
Formal-test inventory	Retained formal Common-5 images/boxes after ontology filtering: PKU 200/924 (100 lineages); MIXED-clean 175/334; SELF 213/585; DeepPCB 150/838; pre-filter counts and excluded-only images are reported in Table S2.
Models and seeds	Primary: YOLOv12m Baseline [22, 34] and Original DynaFocus [35]; confirmation: COCO-pretrained RT-DETR-L [36, 37]; all three detector configurations used seeds 2024, 2025, and 2026.
Primary training	Both primary detectors used Ultralytics 8.3.199, the template-disjoint PKU Common-5 runtime configuration (missing_hole excluded; nc=5), and yolov12m.pt; input size (imgsz) = 960 pixels; 100 epochs; batch=12; SGD (lr0=0.01, lrf=0.01, momentum=0.937, weight decay=5×10−4); cosine LR; patience=30; best.pt selected by Common-5 PKU-validation mAP ₅₀₋₉₅ .
Primary augmentation	HSV=0.015/0.7/0.4; translate=0.10; scale=0.50; fliplr = 0.50; mosaic=1.0 with close_mosaic=10; MixUp=0; no extra Albumentations.
Full-domain evaluation	Images in the retained formal Common-5 population were loaded at native resolution and passed directly to standard Ultralytics letterbox preprocessing at imgsz = 960; no custom pre-resizing was applied; full-image Common-5 GT; minimum confidence threshold=0.001 for precision–recall curve construction; max_det=300; non-maximum suppression (NMS) with an intersection-over-union (IoU) threshold of 0.65 for the primary detectors and YOLOv12m oracles; RT-DETR-L used its NMS-free validation path; Precision, Recall, mAP ₅₀ , and mAP ₅₀₋₉₅ as 3-seed mean ± standard deviation (SD).
Evaluation-chain controls	GT-as-prediction through the same full-domain metric path; three-seed YOLOv12m target oracles used non-test train/validation data only and the unchanged formal test split; cross-split image-key and exact SHA-256 overlaps were zero, and all pHash≤6 candidates were manually reviewed with no confirmed image-level near duplicate.
Fixed target library	600 selected targets: 200/domain, 40/class/domain; 567 unique base images.
Fixed-instance PKU anchor	200 selected targets: 40/class × 5 classes; 100 lineages; template-disjoint from model development.
Paired imaging protocol	Brightness, contrast, and Gaussian blur; reference + 8 levels/factor; 25 unique one-factor-at-a-time conditions.
Selected-target evaluation	Primary: imgsz=960; conf=0.25; match IoU=0.50; NMS IoU=0.70; max_det=300. IO sensitivity: imgsz=640/800/960; conf=0.01/0.05/0.10/0.25/0.50; match IoU=0.30/0.50/0.70.
Pairing and inference	Fixed-instance pairing unit: sample_id; paired-bootstrap resampling unit: lineage_cluster_id; three frozen seed-specific checkpoints.
Auxiliary diagnostics	Annotation validation: 200 images (50/domain), six-class inter-reviewer audit, independent Reviewer C Common-5 validation, prediction-blind final adjudication, and frozen-prediction Recall rescoring; size bins: pooled selected-target area-ratio tertiles (Small/Medium/Large).
Metric boundary	Selected-target analyses do not compute global Precision, F1, AP, or mAP from incomplete image-level GT.

Two evaluation-chain positive controls were performed after the source-only protocol was locked. First, canonical ground-truth boxes were converted to confidence-1 pseudo-predictions and passed through the same prediction-coordinate and metric path used for full-domain evaluation. Second, three-seed YOLOv12m oracles were trained from scratch (pretrained=False) on the locked Common-5 train/validation splits (MIXED-clean: 1305/261 images; SELF: 976/210; DeepPCB: 1122/224) using the 100-epoch, imgsz=960, batch=12 SGD/cosine protocol and augmentations listed for the primary YOLOv12m runs. Each seed’s saved best.pt was evaluated only on the unchanged formal Common-5 test set. Across the three target-oracle splits, cross-split image-key and exact SHA-256 overlaps were zero, and all 235 pHash candidates at the locked distance threshold (≤6) were manually reviewed with no confirmed image-level near duplicate. This audit was limited to image-content integrity and did not establish parent/source, board-template, or physical-PCB disjointness. These controls test shared label-to-metric integrity and target-domain learnability without influencing any source-trained model.

4.2 Annotation Audit and Ground-Truth Validation

Annotation consistency was assessed on a locked 200-image subset, with 50 images from each domain [17, 18]. Samples were selected with a fixed seed (2024) using a coverage-oriented procedure that prioritized defect classes, Small/Medium/Large scales, distinct parent groups, and image-quality conditions including low contrast and low

sharpness. In the six-class canonical audit, which additionally retained ‘missing_hole’ for annotation review, two reviewers inspected each image with the existing annotation visible and independently retained or revised classes and boxes; their outputs were stored separately and used to quantify inter-reviewer agreement. Reviewer C, who was independent of model development and did not participate in the A/B visible-label audit, had prior PCB-defect annotation experience with the formal Common-5 ontology and independently assessed the same images without access to the existing labels or detector predictions.

For the inter-reviewer audit, same-class matches with $\text{IoU} \geq 0.75$ were considered consistent; $0.50 \leq \text{IoU} < 0.75$ indicated boundary disagreement and $\text{IoU} < 0.50$ severe boundary disagreement. Class mismatches, missing annotations, split/merge cases, and uncertain reviews were recorded separately. A/B disputes were adjudicated while preserving all annotation versions. Reviewer C independently assessed the Original Common-5 labels. Same-class pairs with $\text{IoU} \geq 0.50$ retained Original geometry; previously unmatched same-class pairs were auto-reconciled only when $\text{IoU} \geq 0.10$, normalized center distance ≤ 0.05 , the C-to-Original area ratio was 0.5–2.0, and the match was mutually unique. Class changes, unresolved one-sided boxes, ambiguous candidates, and split/merge-like cases underwent final adjudication by a separate senior researcher independent of model development and blinded to detector predictions; the final Common-5 GT was locked before rescoreing.

Image-level exact agreement, box-pair agreement, class agreement, mean IoU, and Cohen’s κ characterize inter-reviewer consistency. Under the locked reconciliation protocol, independent Common-5 validation re-evaluated label presence and class assignments while preserving Original geometry for reconciled same-class matches; Recall from identical frozen predictions was then recomputed against the original and adjudicated labels on the same images. Neither analysis is interpreted as a causal decomposition.

4.3 Fixed-Instance Controlled Imaging Evaluation

Controlled imaging responses were evaluated using a fixed set of target observations constructed before perturbation generation. From each of MIXED-clean, SELF, and DeepPCB, 40 observations were selected for each of the five canonical classes, yielding 200 observations per domain and 600 in total. Selection used a fixed seed of 2026 and did not access detector predictions. The final set originated from 567 unique base images.

Sampling prioritized distinct defect lineages when sufficient lineage capacity was available. Where fewer than 40 independent lineages existed within a domain–class stratum, distinct fixed image observations were used instead, while repeated lineage membership remained explicitly recorded. This preserved balanced class coverage without treating dependent observations as independent samples.

Each observation used the unmodified RGB reference I_0 and three one-factor-at-a-time perturbation families. For $a \in \{0.95, 0.90, 0.85, 0.80, 0.75, 0.70, 0.60, 0.50\}$, brightness used $I_B = \text{clip}(\text{round}(aI_0), 0, 255)$, while contrast used $I_C = \text{clip}(\text{round}(\mu + a(I_0 - \mu)), 0, 255)$, where μ is the image-mean luminance from RGB weights 0.299/0.587/0.114. Gaussian blur used $\sigma = 0.5\text{--}4.0$ native-image pixels with $3 \times 3\text{--}17 \times 17$ kernels and REFLECT_101 padding before letterbox preprocessing; σ was not normalized to model-input pixels across domains. All variants were generated directly from I_0 as uint8 images without geometric transforms, yielding 25 conditions and 15,000 target sample–condition records.

Full-image inference used frozen weights at the primary operating point of $\text{imgsz} = 960$, $\text{conf} = 0.25$, $\text{NMS IoU} = 0.70$, $\text{match IoU} = 0.50$, and $\text{max_det} = 300$. I_0 sensitivity comprised nine unique one-factor-at-a-time settings spanning $\text{imgsz} = 640/800/960$, $\text{conf} = 0.01/0.05/0.10/0.25/0.50$, and $\text{match IoU} = 0.30/0.50/0.70$, with the shared primary setting counted once; all 25 imaging conditions were evaluated at the primary point. Each `sample_id` uniquely identifies one locked selected-target observation and is retained across its I_0 reference and 24 perturbations; `lineage_cluster_id` is the corresponding locked lineage grouping used for clustered resampling. Because selected-target annotations are incomplete image-level ground truth, this branch reports Recall, true-positive-to-false-negative (TP-to-FN) and false-negative-to-true-positive (FN-to-TP) transitions, and paired confidence and localization changes; global Precision, F1, AP, and mAP remain reserved for full-domain evaluation.

A template-disjoint PKU source anchor was selected from the locked Common-5 test pool using seed 2026 without access to detector predictions (40/class; 200 observations; 100 lineages) and evaluated under the same protocol to distinguish perturbation insensitivity from floor effects. It provides a source-domain reference for local sensitivity without converting imaging effects into a causal fraction of the source-to-target gap.

4.4 Scale, Domain-Characteristic, and Error Analysis

To examine differences beyond the controlled perturbations, we characterized defect scale, domain statistics, and selected-target failure states separately. Scale and domain-characteristic analyses were post hoc, whereas the

primary error analysis reused the frozen weights and locked targets at the 960-pixel operating point; no training, target reselection, or perturbation regeneration was performed.

The primary annotation-space scale descriptor was the normalized bounding-box area in Eq. (1), where w and h denote the defect-box width and height, and W and H denote the corresponding image dimensions.

$$r_a = \frac{wh}{WH}, \quad (1)$$

The Small, Medium, and Large categories used tertile boundaries predefined from the pooled selected-target library, not thresholds fitted to the four-domain characterization data. The fixed cut points were $Q_{33} = 0.0024137125$ and $Q_{67} = 0.0053328451$, and the three scale categories were assigned according to Eq. (2).

$$S(r_a) = \begin{cases} \text{Small}, & r_a \leq Q_{33}, \\ \text{Medium}, & Q_{33} < r_a \leq Q_{67}, \\ \text{Large}, & r_a > Q_{67}. \end{cases} \quad (2)$$

Because these thresholds were fixed before examining detector outcomes, subsequent scale-conditioned comparisons do not use performance-dependent binning. These bins were used for the selected-target and complete formal-test analyses. A separate robustness check recomputed input-space box area and short-side pixels after standard letterbox resizing at $\text{imgsz} = 640$ and 960 , using tertiles fixed from the pooled candidate pool used before fixed-target selection, without detector outcomes.

Domain characterization was then performed over the complete formal test data from PKU, MIXED-clean, SELF, and DeepPCB. Object geometry and composition were computed only from the five canonical classes defined in Section 3.1. For domain characterization only, native boxes excluded from the common ontology were omitted from canonical object statistics but retained in the masking set. Each native GT box was expanded by $m = \max(\text{round}(0.05l_b), 2)$ pixels, where l_b is its longer side; masked pixels were excluded from background descriptors. The valid-background ratio exceeded 0.50 for all images (minimum 0.890).

Three descriptor groups were extracted. Per-image visual descriptors comprised grayscale mean and SD, mean local contrast, edge density, Laplacian variance, entropy, mean HSV saturation and value, and gray-level co-occurrence matrix (GLCM) contrast, homogeneity, and energy. Grayscale used an 8-bit RGB-to-L conversion; local contrast used a radius-3 box blur; edge density used FIND_EDGES responses above 32 excluding the outer 2-pixel border; entropy used a 256-bin histogram; and Laplacian variance used a four-neighbor operator. Symmetric 16-level GLCMs pooled horizontal and vertical one-pixel pairs. The same calculations over valid unmasked pixels produced annotation-assisted background descriptors. The second group captures object geometry and density, including canonical object count, objects per megapixel, total and union bounding-box area ratios, mean and median object-area ratio, mean short-side ratio, and mean aspect ratio. The third group describes dataset composition through the proportions of the five canonical defect classes and the fixed Small, Medium, and Large size categories. This separation allows differences in overall image formation, PCB background appearance, and defect composition to be examined without collapsing them into a single domain label. For visualization only, the fixed 14-descriptor display subset is encoded in the figure-generation script rather than selected from the data; the combined principal component analysis (PCA) uses the full locked descriptor set.

For each descriptor, domain-level distributions were summarized independently before comparing each target domain with PKU. Target-to-source differences were evaluated using Eq. (3), where $\tilde{x}_d$ and $\tilde{x}_{\text{PKU}}$ denote the target-domain and PKU medians, respectively, for $d \in \{\text{MIXED-clean}, \text{SELF}, \text{DeepPCB}\}$.

$$\Delta\tilde{x}_d = \tilde{x}_d - \tilde{x}_{\text{PKU}}, \quad (3)$$

Uncertainty was estimated using 2000 clustered bootstrap repetitions with a domain-specific analysis_cluster_id: PKU original/rotation pairs, SELF tile/augmentation variants, and MIXED-clean Roboflow exports were grouped by their source root, while DeepPCB retained the full image stem. Standardized mean differences (SMDs) were retained as complementary effect-size descriptors. These statistics were used to characterize the magnitude and direction of domain differences, not to infer that any individual descriptor causally determines detector degradation.

Because the descriptors are correlated, PCA provided a lower-dimensional view of domain structure. Two primary spaces were fitted: visual-plus-background and combined, which additionally included geometry, scale composition, and class composition; a separate label-free visual-only space was used for robustness analysis. The fitting procedure retained one deterministic representative per cluster and sampled equal cluster counts across domains with seed 20260817. Features with >25% missingness or zero variance in the fitting set were excluded;

remaining values were median-imputed and standardized from fitting-set statistics. Five principal components (PCs) were retained, and all formal-test images were projected with the fitted transformation.

Domain separation was summarized by Euclidean centroid distances in five-dimensional PC space and interpreted descriptively, not as causal or proportional measures of detection loss. For robustness, all three spaces were recomputed after isotropic long-side resizing to 640 px without crop or padding; 500 cluster-level PCA refits and 2000 centroid bootstraps assessed stability.

For each selected I_0 target, deterministic one-to-one matching assigned mutually exclusive states in this order: same-class TP at $\text{IoU} \geq$ the match threshold; residual wrong-class overlap at $\text{IoU} \geq 0.10$ as E3; residual same-class overlap with $0 < \text{IoU} <$ the match threshold as E2; otherwise E1. The primary selected-target analysis used $\text{imgsz} = 960$, $\text{conf} = 0.25$, and $\text{match IoU} = 0.50$. Candidate diagnostics used predictions captured at $\text{conf} = 0.001$ to record maximum-IoU candidates, while the I_0 sensitivity analysis also reported class-agnostic Recall; neither changed the formal states. From the same 960-pixel paired outputs, illustrative cases included TP→E1 blur/contrast transitions and the largest confidence/IoU drops among paired TP→TP cases.

Error summaries were stratified jointly by domain, canonical class, and the fixed size categories. To avoid drawing strong conclusions from small or highly dependent cells, a stratum was considered interpretable only when it contained at least 10 selected targets and at least 10 independent lineages. All remaining cells were retained for completeness but marked as descriptive only. Within interpretable strata, the dominant I_0 failure state was defined from the relative frequencies of these mutually exclusive states; exact ties were retained as joint dominant states rather than being resolved arbitrarily.

The main selected-target analysis focuses on I_0 before any synthetic imaging perturbation is applied. The same taxonomy was nevertheless propagated across all 25 imaging conditions, and transitions from I_0 to each of the 24 perturbed conditions were retained as supplementary evidence of how controlled brightness, contrast, or blur can alter an existing failure state. These transitions are interpreted as perturbation-specific state changes and are not used to estimate a causal contribution of content or imaging factors to the total source-to-target performance gap.

4.5 Statistical Analysis

Statistical analysis was matched to ground-truth completeness. For full-domain evaluation, 0.001 was a fixed confidence floor for precision–recall curve construction. Reported Precision and Recall are Ultralytics mean class-wise summaries at each evaluation set’s F1-maximizing curve index, not common-threshold deployment metrics; mAP_{50} and mAP_{50-95} use the full precision–recall curves. Each metric was computed per training seed and summarized as mean $\pm$ standard deviation across three seeds. For source-only transfer rows, mAP_{50-95} retention was calculated within each seed $s \in \{2024, 2025, 2026\}$ using Eq. (4) and then summarized across seeds.

$$\text{Retention}_{d,s} = \frac{\text{mAP}_{50-95,d,s}}{\text{mAP}_{50-95,\text{PKU},s}} \times 100\%. \quad (4)$$

A supplementary lineage-aware analysis evaluated whether counting both original and rotated PKU test images affected the source estimate or the resulting retention conclusions (Table S6).

For the fixed-instance imaging experiments, selected-target Recall under condition c , denoted R_c , was defined as the proportion of selected targets successfully detected using the fixed confidence and matching criteria. Each perturbed condition was paired with the same unmodified reference I_0 .

$$\Delta R_c = R_c - R_{I_0}, \quad D_c = R_{I_0} - R_c. \quad (5)$$

The signed Recall change and the corresponding degradation were defined in Eq. (5). Accordingly, positive ΔR_c indicates improvement, whereas positive D_c indicates degradation. For imaging sensitivity, mean I_0 Recall ≤ 0.005 defined the primary floor; otherwise, onset was the first level with mean Recall degradation ≥ 0.05 and positive degradation in all three seeds; threshold sensitivity around this criterion is reported in Table S7. Pairing was maintained by sample id .

For each seed, 95% clustered paired-bootstrap confidence intervals (CIs) for D_c were calculated using $\text{lineage_cluster_id}$ as the resampling unit; cross-seed selected-target summaries are reported as mean $\pm$ SD. TP-to-FN and FN-to-TP transitions and paired confidence or localization changes were analyzed under the same pairing structure. For domain-characteristic comparisons, target-minus-PKU median differences were estimated from 2000 with-replacement analysis_cluster_id bootstraps with percentile 95% CIs; SMD was the target-minus-PKU mean difference divided by the pooled sample SD. PCA and centroid-distance analyses remained descriptive and were not treated as hypothesis tests or causal measures of performance loss.

Within the locked 200-image audit subset, annotation analysis combined inter-reviewer agreement with validator-summary Recall changes from identical frozen predictions rescored against the original and adjudicated five-class labels, with the F1-maximizing confidence index determined separately for each label version. Scale- and error-stratified results were eligible for primary interpretation only when their predefined sample and lineage requirements were satisfied; sparse cells were reported descriptively. Effect magnitude, seed consistency, and paired behavior were emphasized; cross-seed SDs quantify training-seed variability, whereas clustered-bootstrap CIs quantify uncertainty under cluster-level resampling for paired imaging effects and domain-characteristic differences. Selected-target analyses were not used to derive global Precision, F1, AP, or mAP, and the diagnostic evidence streams were not combined into additive estimates of the total cross-domain performance gap.

5. Results

5.1 Cross-Domain Performance Degradation and Cross-Architecture Confirmation

At $\text{imgsz} = 960$, all three detector configurations performed substantially better on the template-disjoint PKU split than on MIXED-clean, SELF, and DeepPCB without adaptation (Fig. 2; Table 2). MIXED-clean showed the highest residual detection performance, whereas SELF and DeepPCB remained near zero. Source-domain ranking did not determine target retention, and RT-DETR-L reproduced the severe transfer pattern across a different detector architecture.

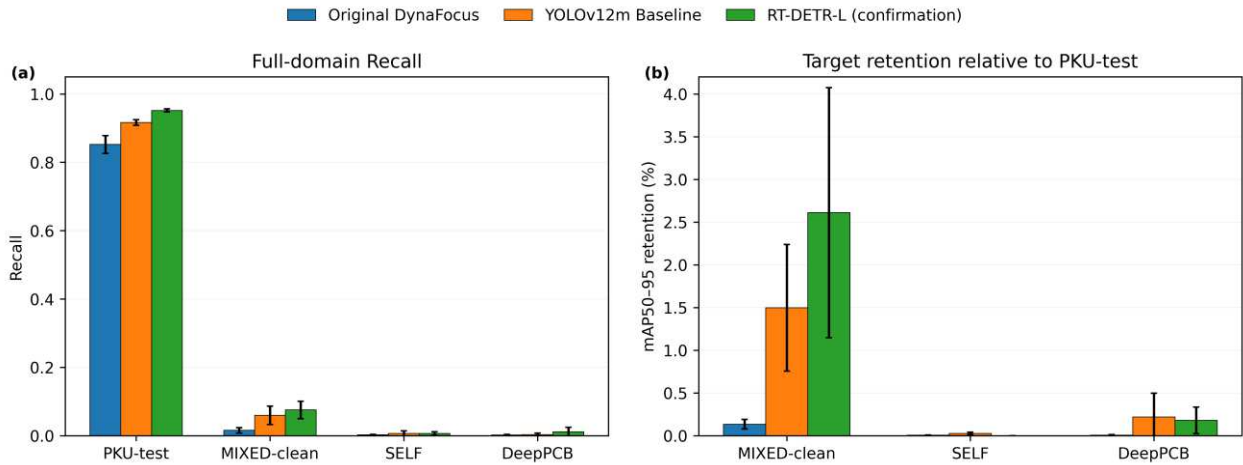


Fig. 2 Cross-domain performance with cross-architecture confirmation at $\text{imgsz} = 960$: (a) full-domain validator-summary Recall and (b) mAP_{50-95} retention relative to PKU. Error bars show ± 1 SD across three training seeds

Using the same template-disjoint Common-5 PKU test split defined in Section 4.1, the rotation-only minus original-only mAP_{50-95} difference was $+0.0094$ (95% CI, 0.0001 – 0.0212) for Original DynaFocus, $+0.0295$ (0.0181 – 0.0410) for YOLOv12m Baseline, and $+0.0160$ (0.0071 – 0.0253) for RT-DETR-L. The severe-transfer conclusion was unchanged: across all source-estimator variants and detectors, the largest bootstrap 95% upper bound for target retention was 2.731% (Table S6).

GT-as-prediction yielded 1.000 Precision/Recall and 0.995 $\text{mAP}_{50}/\text{mAP}_{50-95}$ in all four domains; target-domain YOLOv12m oracles achieved mAP_{50-95} of 0.3811 – 0.9304 across the three targets (Table 2). These controls support the shared label-to-metric path and show that substantially higher within-target performance is attainable under the locked oracle splits, but do not establish parent/template-independent oracle generalization or exclude model-specific preprocessing or decoding errors.

Table 2. Full-domain cross-domain performance and positive controls at imsz = 960

Model / control	Domain	Precision	Recall	mAP ₅₀	mAP ₅₀₋₉₅	mAP retention
Original DynaFocus	PKU-test	0.830 ± 0.070	0.852 ± 0.026	0.877 ± 0.039	0.4466 ± 0.0207	100.00%
Original DynaFocus	MIXED-clean	0.013 ± 0.006	0.016 ± 0.008	0.002565 ± 0.001129	0.000606 ± 0.000271	0.13%
Original DynaFocus	SELF	9.84e−05 ± 8.45e−05	0.002435 ± 0.001489	5.01e−05 ± 4.33e−05	2.21e−05 ± 1.98e−05	<0.01%
Original DynaFocus	DeepPCB	2.16e−04 ± 2.00e−04	0.001815 ± 0.001627	1.09e−04 ± 1.01e−04	2.71e−05 ± 3.01e−05	0.01%
YOLOv12m Baseline	PKU-test	0.906 ± 0.037	0.917 ± 0.008	0.931 ± 0.017	0.4827 ± 0.0122	100.00%
YOLOv12m Baseline	MIXED-clean	0.101 ± 0.020	0.060 ± 0.027	0.0290 ± 0.0144	0.007164 ± 0.003343	1.50%
YOLOv12m Baseline	SELF	0.000899 ± 0.000401	0.006409 ± 0.007830	0.000305 ± 0.000148	0.0001235 ± 0.0000727	0.03%
YOLOv12m Baseline	DeepPCB	0.233 ± 0.397	0.003218 ± 0.004434	0.001967 ± 0.002761	0.001036 ± 0.001293	0.22%
RT-DETR-L (confirmation)	PKU-test	0.943 ± 0.009	0.952 ± 0.004	0.956 ± 0.006	0.4723 ± 0.0107	100.00%
RT-DETR-L (confirmation)	MIXED-clean	0.200 ± 0.056	0.075 ± 0.025	0.0495 ± 0.0322	0.012266 ± 0.006671	2.61%
RT-DETR-L (confirmation)	SELF	5.07e−05 ± 4.39e−05	0.006127 ± 0.005316	3.10e−06 ± 2.89e−06	8.12e−07 ± 7.42e−07	<0.01%
RT-DETR-L (confirmation)	DeepPCB	0.024 ± 0.022	0.011 ± 0.013	0.002246 ± 0.002017	0.000837 ± 0.000729	0.18%
GT-as-prediction	All four domains	1.000	1.000	0.995	0.995	—
YOLOv12m oracle (control)	MIXED-clean	0.865 ± 0.013	0.843 ± 0.014	0.874 ± 0.007	0.3811 ± 0.0047	—
YOLOv12m oracle (control)	SELF	0.995 ± 0.002	0.969 ± 0.003	0.985 ± 0.001	0.9304 ± 0.0029	—
YOLOv12m oracle (control)	DeepPCB	0.840 ± 0.018	0.864 ± 0.005	0.883 ± 0.005	0.6991 ± 0.0065	—

Note: Reported Precision/Recall are validator curve summaries; model metrics are mean ± SD over three seeds, while GT-as-prediction is deterministic across the four formal domains. Source-only retention is computed seed-wise and reported as the cross-seed mean; retention is not defined for positive controls. Oracle train/validation/test memberships passed the bounded image-content integrity audit described in Section 4.1. Source-trained detector rows use the template-disjoint Common-5 checkpoints defined in Section 4.1; mapping and filtering QC are reported in Tables S1–S2.

5.2 Annotation Validation and Recall Impact

With the transfer gap established, the inter-reviewer audit showed high overall agreement but a markedly uneven distribution of disagreement across domains (Fig. 3; Table 3, Panel A). Disagreements were concentrated in MIXED-clean, whereas PKU showed only isolated disagreement and SELF and DeepPCB showed complete image-level agreement between the two reviewers. Even in MIXED-clean, class agreement remained high, indicating localized reviewer disagreement rather than pervasive inter-reviewer semantic inconsistency.

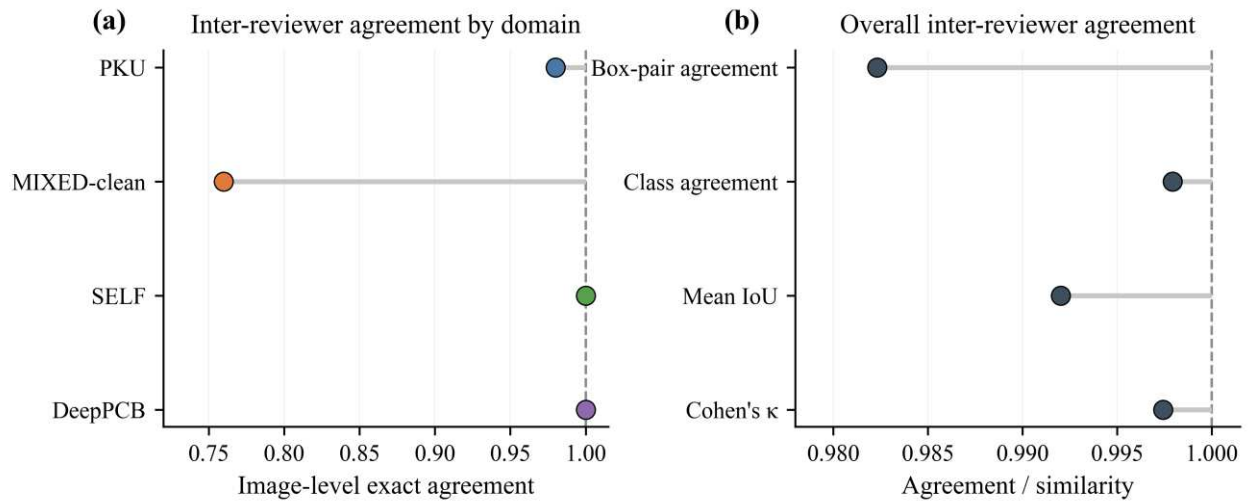


Fig. 3 Six-class inter-reviewer agreement across the four PCB domains: (a) image-level exact agreement and (b) overall object- and class-level agreement

Table 3. Annotation audit, independent Common-5 validation, adjudication, and frozen-prediction Recall rescoring

Panel A. Six-class inter-reviewer agreement

Domain	n	Disputed	Image agr.	Box agr.	Class agr.	Mean IoU	κ
Overall	200	13	93.50%	98.20%	99.80%	0.992	0.997
PKU	50	1	98.00%	99.60%	100.00%	0.999	1.000
MIXED-clean	50	12	76.00%	88.70%	98.60%	0.949	0.983
SELF	50	0	100.00%	100.00%	100.00%	1.000	1.000
DeepPCB	50	0	100.00%	100.00%	100.00%	1.000	1.000

Note: All 13 disputed images were resolved; six used Reviewer A’s annotations and seven used Reviewer B’s, with no unresolved cases.

Panel B. Frozen-prediction Recall rescoring after independent Common-5 validation and adjudication (50 images/domain)

Domain	Model	Recall orig.	Recall adj.	ΔR_{ann} , mean $\pm$ SD
PKU	Original DynaFocus	0.367 ± 0.018	0.319 ± 0.016	-0.048 ± 0.018
MIXED-clean		0.008696 ± 0.000000	0.023 ± 0.016	$+0.014 \pm 0.016$
SELF		0.000430 ± 0.000745	0.000355 ± 0.000614	-0.000075 ± 0.000131
DeepPCB		0.003361 ± 0.003944	0.001887 ± 0.001887	-0.001475 ± 0.002140
PKU	YOLOv12m Baseline	0.398 ± 0.009	0.336 ± 0.008	-0.062 ± 0.002
MIXED-clean		0.053 ± 0.035	0.070 ± 0.030	$+0.018 \pm 0.007$
SELF		0.003175 ± 0.002749	0.005208 ± 0.006505	$+0.002034 \pm 0.005007$
DeepPCB		0.000000 ± 0.000000	0.000000 ± 0.000000	0.000000 ± 0.000000

Note: Reviewer C independently re-evaluated the Original Common-5 labels. Same-class spatial matches retained the Original geometry under the locked reconciliation protocol; only unresolved presence or class discrepancies underwent final adjudication. Rescoring used identical frozen predictions from the locked primary checkpoints on the same 50 images per domain; validator-summary Recall used the F1-maximizing confidence index determined separately for each

label version. Values are mean $\pm$ sample SD across three seeds; $\Delta R_{\text{ann}} = \text{Recall}_{\text{adj.}} - \text{Recall}_{\text{orig.}}$; the table reports the cross-seed mean $\pm$ sample SD of these paired differences.

Panel C. Final Common-5 adjudication changes (50 images/domain)

Domain	Images changed	Original boxes	Final boxes	Added	Removed	Class changes	Box changes
PKU	15/50	197	226	29	0	0	0
MIXED-clean	18/50	109	135	26	0	1	0
SELF	38/50	213	276	63	0	1	0
DeepPCB	35/50	346	407	61	0	24	0
Overall	106/200	865	1044	179	0	26	0

Note: Counts compare the original Common-5 labels with the final adjudicated labels on the locked audit subset. “Images changed” denotes images with at least one final annotation change; Reviewer C screening and intermediate reconciliation differences are not counted as final corrections.

Despite the high A/B agreement, final adjudication changed 106 of 200 audited images, adding 179 boxes and relabeling 26 matched instances without removing or geometrically altering original boxes (Table 3, Panel C). Across the eight model–domain rows, ΔR_{ann} ranged from -0.062 to $+0.018$, with $|\Delta R_{\text{ann}}| \leq 0.018$ in the target domains; adjudicated rescoring therefore preserved the severe source-to-target Recall pattern within the audit subset (Table 3, Panel B). This result does not establish label accuracy for the full formal Common-5 test populations. The following analyses examine controlled imaging sensitivity and broader domain-characteristic differences.

5.3 Controlled Imaging Responses Across Domains

At $\text{imgsz} = 960$, perturbation responses were factor- and domain-dependent and did not uniformly reduce Recall (Fig. 4; Table 4). PKU remained interpretable for both models, whereas MIXED-clean was interpretable only for YOLOv12m Baseline under the primary criterion. For that row, brightness and contrast showed no onset, while Gaussian blur reached onset at native-image $\sigma = 3.5$ px under the present protocol.

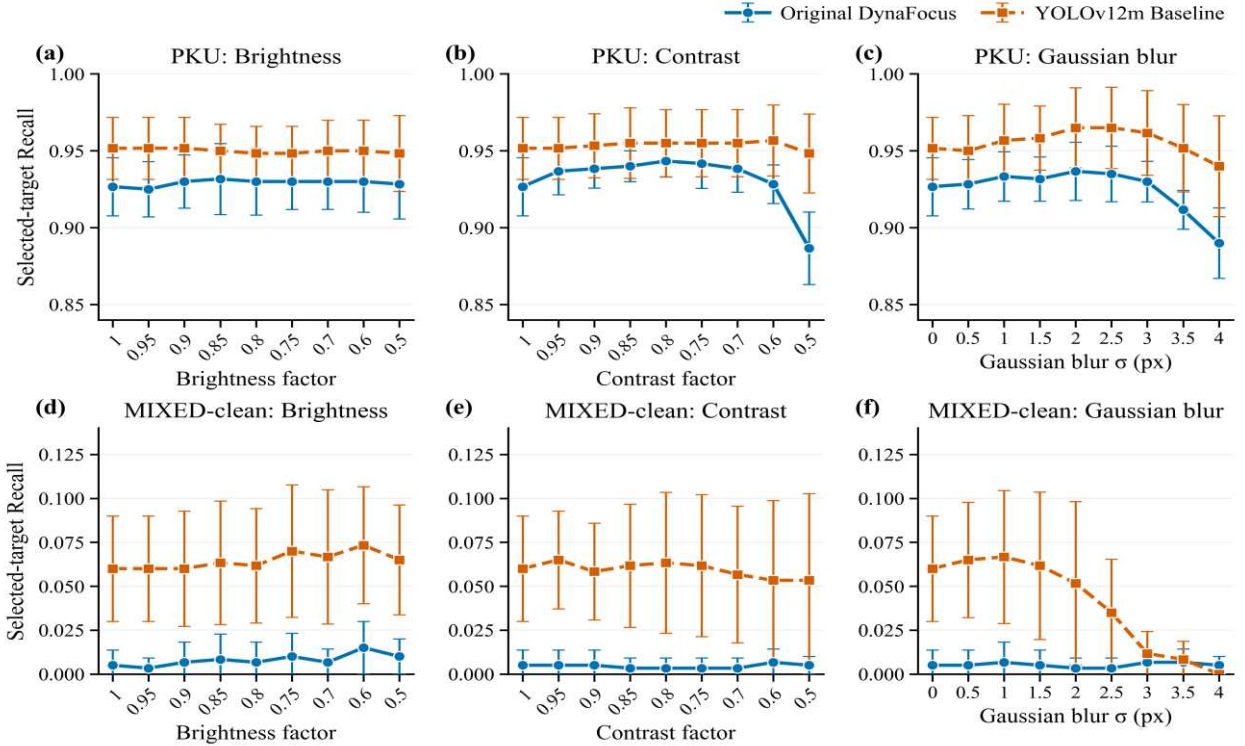


Fig. 4 Selected-target Recall at $\text{imgsz} = 960$ under controlled brightness, contrast, and Gaussian blur: (a–c) PKU and (d–f) MIXED-clean. Error bars show ± 1 SD across three seeds

The PKU anchor retained high I_0 Recall (0.927–0.952), and no tested perturbation reached the primary onset criterion for either model. Responses remained factor-dependent and nonmonotonic; per-seed D_c estimates and 95%

524 clustered paired-bootstrap CIs are reported in Table S3. Threshold sensitivity showed that exact onset assignments
525 varied with the criterion and should therefore be interpreted using the predefined primary criterion (Table S7).

526 Table 4. Selected-target I_0 Recall and imaging-sensitivity onset at $\text{imgsz} = 960$

Model	Domain	I_0 Recall	Status	Brightness onset	Contrast onset	Blur onset
Original DynaFocus	PKU	0.927 ± 0.019	Interpretable	No onset	No onset	No onset
Original DynaFocus	MIXED-clean	0.005 ± 0.009	Floor	—	—	—
Original DynaFocus	SELF	0.000	Floor	—	—	—
Original DynaFocus	DeepPCB	0.000	Floor	—	—	—
YOLOv12m Baseline	PKU	0.952 ± 0.020	Interpretable	No onset	No onset	No onset
YOLOv12m Baseline	MIXED-clean	0.060 ± 0.030	Interpretable	No onset	No onset	3.5 px
YOLOv12m Baseline	SELF	0.000	Floor	—	—	—
YOLOv12m Baseline	DeepPCB	0.0017 ± 0.0029	Floor	—	—	—

527 Note: Recall is reported as mean $\pm$ SD across three seeds; exact-zero rows have zero SD. Floor status was assigned from
528 the unrounded three-seed mean using $\bar{R}_{I_0} \leq 0.005$. Onset is the first perturbation level for which mean Recall
529 degradation is at least 0.05 and degradation is positive in all three seeds.

530 At the primary operating point, SELF and DeepPCB were at or below the I_0 floor; this does not imply imaging
531 robustness under further perturbation. Measurable responses in PKU and YOLOv12m/MIXED-clean showed local
532 imaging sensitivity, while imaging attribution remained unresolved for floor-limited SELF and DeepPCB.

533 5.4 Scale and Domain-Characteristic Differences

534 Beyond the controlled imaging responses, the class-balanced selected-target library exhibited distinct within-
535 library scale profiles (Fig. 5; Table 5): large targets predominated in MIXED-clean, small targets in SELF, and
536 medium targets in DeepPCB. These proportions do not estimate full-domain prevalence. Input-space area and short-
537 side analyses preserved the broad domain pattern at input sizes of 640 and 960 pixels. MIXED-clean nevertheless
538 showed pronounced input-size dependence: selected-target Recall changed from 0.145 ± 0.045 to 0.005 ± 0.009 for
539 DynaFocus and from 0.362 ± 0.060 to 0.060 ± 0.030 for YOLOv12m between 640 and 960. SELF remained at zero,
540 and DeepPCB remained zero or near zero. Size-conditioned results therefore do not support a universal size-
541 vulnerability rule.

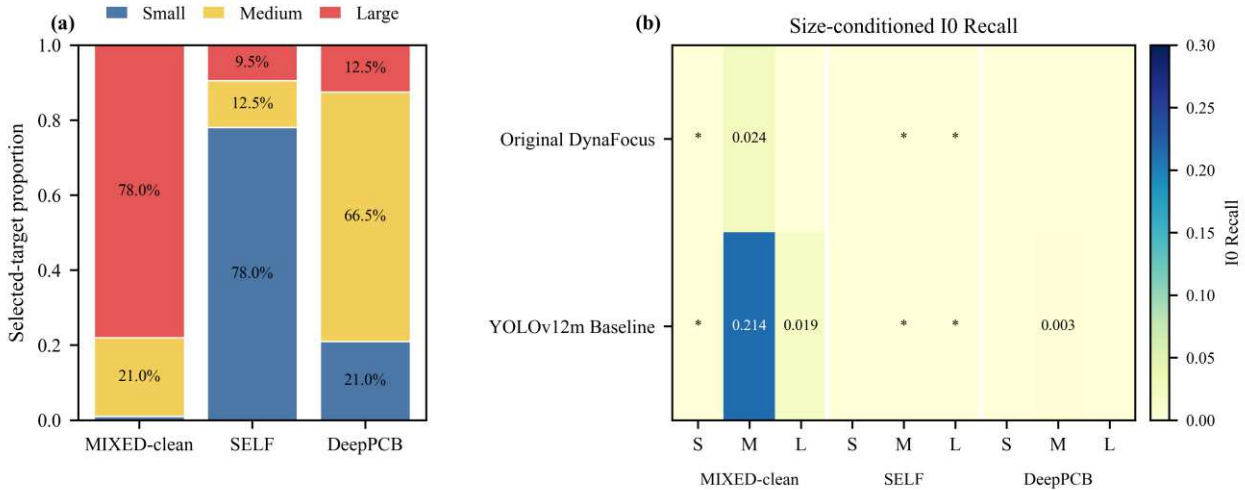


Fig. 5 Selected-target normalized-area scale composition and size-conditioned I_0 Recall at $\text{imgsz} = 960$: (a) Small/Medium/Large composition and (b) size-conditioned Recall. Asterisks mark descriptive-only strata below the predefined interpretability criterion

542 The full-domain characterization confirmed distinct scale, appearance, background, geometry, density, and
543 composition profiles across targets (Fig. 6). DeepPCB differed most strongly in intensity/texture descriptors,
544 MIXED-clean in geometric/scale descriptors, and SELF in contrast/sharpness descriptors; annotation-assisted
545 background descriptors showed corresponding background differences. Target-minus-PKU median differences and
546 95% clustered-bootstrap CIs are reported in Table S4.

547 Table 5. Selected-target normalized-area scale profile and size-conditioned I_0 Recall

Domain	Size	n (share)	Lineages	DynaFocus I_0 Recall	YOLOv12m I_0 Recall	Note
MIXED-clean	Small	2 (1.0%)	2	0.000	0.000	*
MIXED-clean	Medium	42 (21.0%)	38	0.024 ± 0.041	0.214 ± 0.063	
MIXED-clean	Large	156 (78.0%)	91	0.000	0.019 ± 0.023	
SELF	Small	156 (78.0%)	11	0.000	0.000	
SELF	Medium	25 (12.5%)	4	0.000	0.000	*
SELF	Large	19 (9.5%)	3	0.000	0.000	*
DeepPCB	Small	42 (21.0%)	42	0.000	0.000	
DeepPCB	Medium	133 (66.5%)	130	0.000	0.0025 ± 0.0043	
DeepPCB	Large	25 (12.5%)	25	0.000	0.000	

548 Note: Recall is mean $\pm$ SD across three seeds; shares refer to the class-balanced selected-target library, not full-domain
549 prevalence; * indicates a descriptive-only stratum below the predefined interpretability criterion.

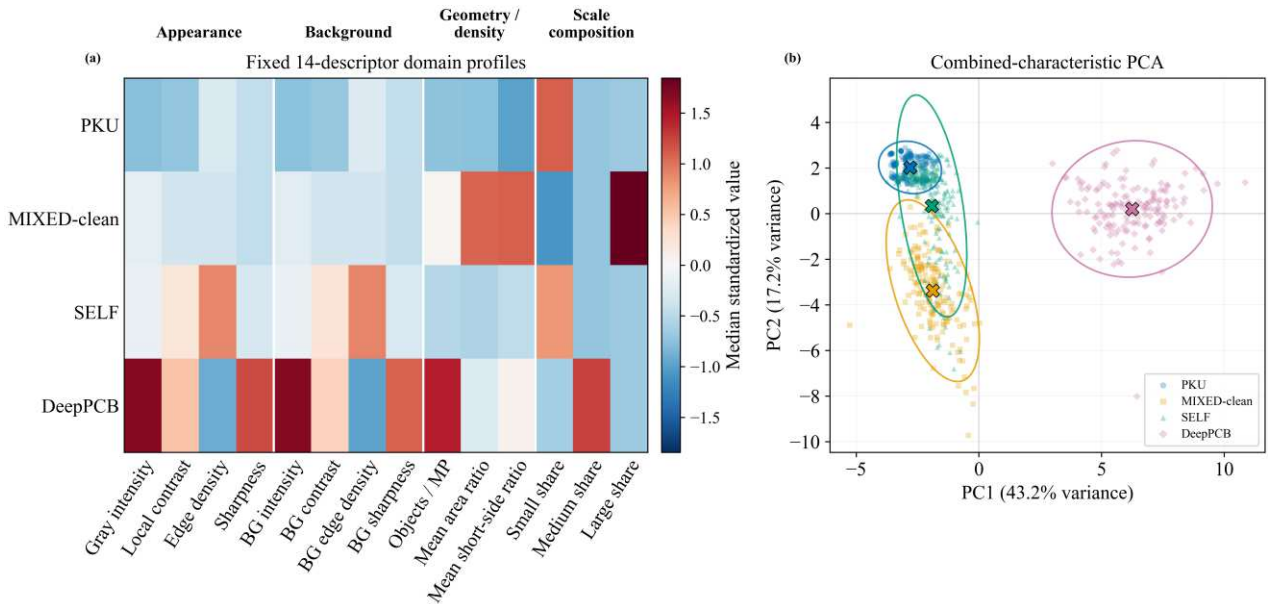


Fig. 6 Full-domain characteristic structure across the four PCB domains: (a) fixed 14-descriptor standardized display subset and (b) combined-characteristic PCA using the full locked descriptor set. Ellipses in (b) are descriptive 95% covariance regions, not confidence intervals

550 Fig. 6(b) shows the combined-characteristic PCA. In the separate label-free visual-only PCA after isotropic long-
551 side normalization to 640 px, the five-dimensional PKU-to-target centroid distances were 2.512 for MIXED-clean,
552 1.247 for SELF, and 6.307 for DeepPCB (Table S5). Spatial normalization changed the distances but did not remove
553 domain separation; the 500-refit and 2000-bootstrap robustness results are reported in Table S5. Adding annotation-
554 assisted background descriptors and then geometry and composition changed the separation magnitude but was not
555 required to obtain visual-only separation. These descriptive differences coexist with the degradation established in
556 Section 5.1 but do not establish causal contributions.

557 5.5 Selected-Target Failure States

558 To characterize selected-target detector failure directly, E1 was the most frequent I_0 state in all three targets at
559 the primary operating point (Fig. 7(a)). Every interpretable MIXED-clean and DeepPCB domain–class–size stratum
560 was E1-dominant for both models, whereas no SELF stratum met the prespecified criterion (Table 6). E1 remained
561 most frequent across all nine I_0 sensitivity settings for every model–target pair (Table S8).

562 The domain–class–size analysis further showed that the selected-target failure pattern cannot be reduced to a
563 universal defect-size rule. Interpretable strata exhibited domain- and class-conditioned variation, whereas strata with
564 insufficient independent lineage support were retained only as descriptive evidence. At conf = 0.01, mean class-
565 agnostic Recall was 0.077–0.132 on MIXED-clean and ≤ 0.022 on SELF/DeepPCB. At the 0.001 capture threshold,

mean seed-wise shares of primary E1 cases with no canonical-overlap candidate were 42.2–44.2%, 63.4–72.6%, and 73.1–91.7% for MIXED-clean, SELF, and DeepPCB, respectively (Table S8). The stratified results therefore refine the scale analysis in Section 5.4 by distinguishing differences in defect composition from the actual form of detector failure.

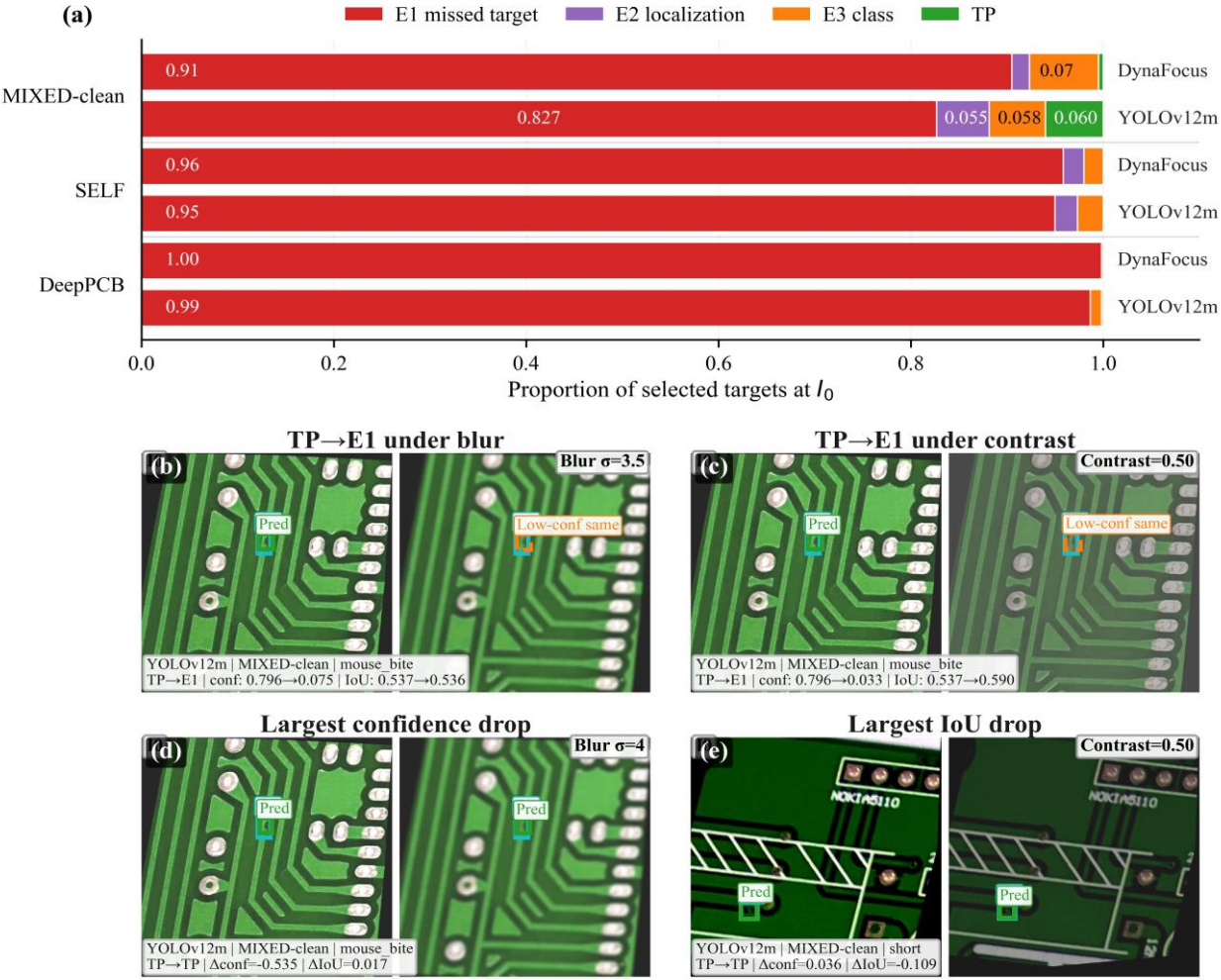


Fig. 7 Selected-target failure states and illustrative paired cases at an input size of 960 pixels, a confidence threshold of 0.25, and a matching IoU threshold of 0.50. (a) Aggregate I_0 state proportions; (b–c) TP → E1 transitions under blur and contrast. The orange box shows the highest-overlap same-class candidate from the diagnostic capture at conf = 0.001 when its confidence falls below the formal 0.25 threshold. (d–e) Largest confidence and IoU decreases among paired TP → TP cases. E1, E2, and E3 denote missed-target, localization-failure, and classification-confusion states, respectively

Table 6. Dominant I_0 failure states at imgsiz = 960 across interpretable domain–class–size strata

Domain	Model	Interpretable strata	E1 missed-target dominant	E2 localization dominant	E3 class dominant	Tie
MIXED-clean	Original DynaFocus	7	7 (100.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
MIXED-clean	YOLOv12m Baseline	7	7 (100.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
SELF	Original DynaFocus	0	—	—	—	—
SELF	YOLOv12m Baseline	0	—	—	—	—
DeepPCB	Original DynaFocus	8	8 (100.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
DeepPCB	YOLOv12m Baseline	8	8 (100.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)

Note: E1, E2, and E3 denote missed-target, localization-failure, and classification-confusion states, respectively; Tie indicates equal dominant rates.

Illustrative paired cases show how controlled perturbations can alter detection outcomes for targets detected under I_0 . Blur or contrast changes can convert an initially detected defect into a miss, while other cases exhibit substantial reductions in confidence or localization overlap. These examples illustrate perturbation-specific transitions but do not by themselves explain the target-domain failures under the reference condition.

6. Discussion

6.1 Integrated Diagnostic Interpretation of the Cross-Domain Gap

No examined diagnostic branch alone provided a sufficient account of the cross-domain gap. Severe degradation appeared in both primary detectors and RT-DETR-L, while the positive controls supported the shared label-to-metric path and substantially higher within-target performance under the locked oracle splits. Residual performance still differed across MIXED-clean, SELF, and DeepPCB, so transfer difficulty remained domain dependent.

Inter-reviewer disagreement was concentrated in MIXED-clean, while independent five-class validation and adjudicated Recall rescoring within the 200-image audit subset did not change the severe source-to-target Recall pattern. Controlled imaging changes affected measurable detections in PKU and YOLOv12m/MIXED-clean, showing local sensitivity to the tested image-formation changes [14, 15]. SELF and DeepPCB were at the primary I_0 detection floor, so the absence of further Recall loss cannot be interpreted as robustness; imaging attribution therefore remains unresolved in these domains.

The domain-characterization results further describe this heterogeneity. The target datasets differed from PKU in image appearance, PCB background, defect geometry, density, scale distribution, and class composition. Input-space scale checks preserved the broad domain ordering, while selected-target results did not support a universal size-dependent failure rule. Residual MIXED-clean detection was input-size sensitive, but missed-target failures remained dominant across all nine I_0 sensitivity settings; the 0.001 candidate audit showed higher no-overlap shares in SELF/DeepPCB than in MIXED-clean (Table S8).

Taken descriptively, the severe degradation co-occurred with differences in image appearance [10], PCB background, defect geometry, scale, and dataset composition [8, 9, 23]. These observations support a multidimensional diagnosis of the examined transfers but do not identify causal contributions from individual or combined domain differences.

6.2 Implications for Cross-Domain PCB Inspection

First, in the present PKU-source setting, higher source-domain performance did not imply target-domain readiness: all three detector configurations performed substantially better on PKU than on the target domains. This result supports direct target-domain evaluation rather than reliance on source-domain accuracy alone [8–10].

Second, robustness assessment should extend beyond conventional image-quality perturbations [14, 15]. Brightness, contrast, and blur produced interpretable responses in PKU and YOLOv12m/MIXED-clean, whereas the I_0 floor in SELF and DeepPCB prevented attribution of their reference-condition failures to these perturbations. A detector that remains stable under a limited corruption test may therefore still be vulnerable to changes in PCB background, defect geometry, scale distribution, density, or dataset composition [8, 9, 30]. Controlled perturbation tests should therefore be interpreted alongside characterization of the target-domain data distribution.

Third, annotation quality remains important for dataset quality assurance [17, 38], but within the 200-image audit subset, adjudicated Recall rescoring did not change the severe transfer Recall pattern. Annotation auditing is therefore useful for identifying evaluation uncertainty and domain-specific labeling problems, while model robustness must be addressed separately when substantial visual and compositional mismatch remains.

Finally, missed-target dominance persisted across all nine I_0 sensitivity settings despite partial MIXED-clean recovery at lower confidence thresholds and smaller input sizes (Table S8); operating-point adjustment did not resolve this selected-target failure pattern. For the examined PKU-source transfers, model acceptance should therefore include target-domain validation, data characterization, and selected-target failure-state analysis before adaptation.

6.3 Limitations

Several limitations define the scope of the present diagnosis. First, the four datasets represent heterogeneous PCB inspection domains rather than a controlled factorial acquisition across four physical production lines. Differences in acquisition appearance, PCB background, defect geometry, scale, composition, and annotation characteristics therefore coexist across domains. The observed associations cannot be interpreted as independent causal effects, and the present study cannot assign a quantitative fraction of the source-to-target performance gap to

any individual observed difference or diagnostic branch. In addition, PKU is the sole source domain, and RT-DETR-L provides one cross-architecture confirmation; the empirical conclusions are therefore restricted to transfer from PKU to MIXED-clean, SELF, and DeepPCB and should not be generalized to other source–target directions or detector architectures. All three observational targets fell within a severe transfer-degradation regime; the study therefore does not establish a calibrated severity spectrum or transfer thresholds.

Second, the controlled imaging experiments isolate only brightness, contrast, and Gaussian blur. These perturbations do not represent the full range of industrial imaging variation. Interpretation is further constrained because SELF and DeepPCB were already at or below the primary I_0 floor, residual MIXED-clean selected-target detection was input-size sensitive, and full-domain metrics were evaluated only at $\text{imgsz} = 960$. Additional perturbations therefore provide limited evidence about intrinsic imaging sensitivity in floor-limited domains.

Finally, several diagnostic branches deliberately trade breadth for traceability. The annotation analyses, including independent five-class validation and Recall rescoring, were limited to the fixed 200-image audit subset rather than the full formal Common-5 test populations. The selected-target experiments use incomplete image-level ground truth and support Recall, paired transitions, confidence, localization, and GT-associated failure-state analyses, but not background false-positive or duplicate-detection prevalence, global Precision, F1, AP, or mAP. Some domain–class–size strata also lacked sufficient independent lineage support for strong interpretation. Likewise, the handcrafted domain descriptors and PCA characterize measurable differences in the selected feature space rather than providing a complete representation of domain shift. These constraints limit the granularity of attribution, but they preserve a clear distinction between observed cross-domain degradation and the diagnostic evidence used to investigate it.

7. Conclusion

This study systematically diagnosed the degradation of PKU-trained PCB defect detectors when transferred to MIXED-clean, SELF, and DeepPCB. At $\text{imgsz} = 960$, the two primary detectors and RT-DETR-L showed severe transfer degradation; GT-as-prediction yielded 1.000 Precision/Recall and 0.995 $\text{mAP}_{50}/\text{mAP}_{50-95}$, while target-domain oracle mAP_{50-95} ranged from 0.3811 to 0.9304. These controls support the shared label-to-metric path and show substantially higher within-target performance under the locked oracle splits, without establishing target-oracle parent/template independence or validating model-specific preprocessing or decoding. No examined diagnostic branch alone provided a sufficient account of this degradation. Within the locked 200-image audit subset, independent five-class validation and adjudicated Recall rescoring did not alter the severe source-to-target Recall pattern, while controlled brightness, contrast, and blur revealed local sensitivity where reference-condition detectability remained measurable; imaging attribution remained unresolved for floor-limited target domains.

The target domains also differed from the source across multiple characteristics, including image appearance, PCB background, defect geometry, scale distribution, and dataset composition. These differences were accompanied by domain- and class-conditioned selected-target patterns, with missed-target states remaining dominant across all nine I_0 sensitivity settings. The overall evidence therefore shows that the observed degradation co-occurred with multiple measurable domain mismatches.

Within this PKU-source study, the results support direct target-domain evaluation before deployment and a diagnostic workflow combining full-domain performance evaluation, annotation auditing, controlled imaging tests, domain characterization, and selected-target failure-state analysis. Because the datasets do not constitute a controlled factorial acquisition, these evidence streams provide diagnostic evidence rather than additive causal contributions. Validation with additional source domains and target domains spanning milder transfer regimes is required before broader generalization.

Statements and Declarations

Funding: No funding was received for conducting this study.

Competing interests: The authors have no relevant financial or non-financial interests to disclose.

Ethics approval: This study did not involve human participants, human data, or animals; therefore, ethics approval was not required.

Data availability: The derived data, dataset manifests, experimental protocols, audit records, evaluation outputs, statistical results, and source data underlying Tables S1–S8 are available in the accompanying Mendeley Data repository at <https://doi.org/10.17632/p9ts8kz6gd.1>. The SELF images, annotations, and split files are available in Mendeley Data, Version 5, at <https://doi.org/10.17632/tjhpnhzmrp.5>. The MIXED-PCB source used for MIXED-

clean is available from its original publication and Mendeley Data, Version 1, at <https://doi.org/10.17632/fj4krvmrr5.1>. PKU and DeepPCB are available from their original cited sources.

Code availability: The custom scripts, notebooks, model configurations, evaluation configurations, and software-environment specifications used for data processing, training, evaluation, statistical analysis, and figure generation are available in the accompanying Mendeley Data repository at <https://doi.org/10.17632/p9ts8kz6gd.1>.

References

1. Ling, Q., Isa, N.A.M.: Printed Circuit Board Defect Detection Methods Based on Image Processing, Machine Learning and Deep Learning: A Survey. *IEEE Access*. 11, 15921–15944 (2023). <https://doi.org/10.1109/ACCESS.2023.3245093>
2. Zhou, Y., Yuan, M., Zhang, J., Ding, G., Qin, S.: Review of vision-based defect detection research and its perspectives for printed circuit board. *J. Manuf. Syst.* 70, 557–578 (2023). <https://doi.org/10.1016/j.jmsy.2023.08.019>
3. Chen, X., Wu, Y., He, X., Ming, W.: A Comprehensive Review of Deep Learning-Based PCB Defect Detection. *IEEE Access*. 11, 139017–139038 (2023). <https://doi.org/10.1109/ACCESS.2023.3339561>
4. Zhang, L., Chen, J., Chen, J., Wen, Z., Zhou, X.: LDD-Net: Lightweight printed circuit board defect detection network fusing multi-scale features. *Eng. Appl. Artif. Intell.* 129, 107628 (2024). <https://doi.org/10.1016/j.engappai.2023.107628>
5. Xudong, S., Xiumin, S.: SNF-YOLOv8: A Lightweight PCB Defect Detection Algorithm base on Multiscale Feature Fusion and Attention Scale Sequence Fusion. *Journal of Electronic Testing: Theory and Applications (JETTA)*. 41, 561–573 (2025). <https://doi.org/10.1007/s10836-025-06189-z>
6. Wang, H., Xie, J., Xu, X., Zheng, Z.: Few-Shot PCB Surface Defect Detection Based on Feature Enhancement and Multi-Scale Fusion. *IEEE Access*. 10, 129911–129924 (2022). <https://doi.org/10.1109/ACCESS.2022.3228392>
7. Wang, J., Ma, L., Li, Z., Cao, Y., Zhang, H.: SSHP-YOLO: A High Precision Printed Circuit Board (PCB) Defect Detection Algorithm with a Small Sample. *Electronics (Basel)*. 14, 217 (2025). <https://doi.org/10.3390/electronics14020217>
8. Yan, Z., Hao, R., Huang, B., Zhu, L., Pan, H.: A Domain Incremental Learning Framework for PCB Continuous Defect Detection. *IEEE Trans. Instrum. Meas.* 74, 1–13 (2025). <https://doi.org/10.1109/TIM.2025.3550210>
9. Yin, X., Wang, S., Wang, Y., Wang, P., Liu, Y., Shen, T., Qiao, H.: Zero-shot printed circuit board defect detection via optical flow and reconstruction guidance. *Comput. Ind.* 173, 104355 (2025). <https://doi.org/10.1016/j.compind.2025.104355>
10. Nguyen, H.D., Cheng, D., Wang, X., Shi, Y., Wen, B.: Data-Efficient Deep Learning for Printed Circuit Board Defect Detection Using X-Ray Images. *IEEE Trans. Instrum. Meas.* 74, 1–14 (2025). <https://doi.org/10.1109/TIM.2025.3529089>
11. Vidit, V., Salzmann, M.: Attention-based domain adaptation for single-stage detectors. *Mach. Vis. Appl.* 33, 65 (2022). <https://doi.org/10.1007/s00138-022-01320-y>
12. Deng, Z., Kong, Q., Akira, N., Yoshinaga, T.: Hierarchical contrastive adaptation for cross-domain object detection. *Mach. Vis. Appl.* 33, 62 (2022). <https://doi.org/10.1007/s00138-022-01317-7>
13. Ma, S., Song, K., Niu, M., Tian, H., Yan, Y.: Cross-scale fusion and domain adversarial network for generalizable rail surface defect segmentation on unseen datasets. *J. Intell. Manuf.* 35, 367–386 (2024). <https://doi.org/10.1007/s10845-022-02051-7>
14. Rodríguez-Rodríguez, J.A., López-Rubio, E., Ángel-Ruiz, J.A., Molina-Cabello, M.A.: The Impact of Noise and Brightness on Object Detection Methods. *Sensors*. 24, 821 (2024). <https://doi.org/10.3390/s24030821>
15. Park, J.H., Kim, Y.S., Seo, H., Cho, Y.J.: Analysis of Training Deep Learning Models for PCB Defect Detection. *Sensors*. 23, 2766 (2023). <https://doi.org/10.3390/s23052766>
16. Honzátko, D., Türetken, E., Bigdeli, S.A., Dunbar, L.A., Fua, P.: Defect segmentation for multi-illumination quality control systems. *Mach. Vis. Appl.* 32, 118 (2021). <https://doi.org/10.1007/s00138-021-01244-z>
17. Nou, S., Lee, J.S., Ohyama, N., Obi, T.: The improvement of ground truth annotation in public datasets for human detection. *Mach. Vis. Appl.* 35, 49 (2024). <https://doi.org/10.1007/s00138-024-01527-1>
18. Freire, A., Leandro, L.H., de Andrade, J.V.R., Azevedo, G.O.A., Fernandes, B.J.T.: Beyond clean data: Exploring the effects of label noise on object detection performance. *Knowl. Based. Syst.* 304, 112544 (2024). <https://doi.org/10.1016/j.knosys.2024.112544>
19. Yao, N., Zhao, Y., Kong, S.G., Guo, Y.: PCB defect detection with self-supervised learning of local image patches. *Measurement*. 222, 113611 (2023). <https://doi.org/10.1016/j.measurement.2023.113611>
20. Liu, T., Cao, G.Z., He, Z., Xie, S.: Refined Defect Detector With Deformable Transformer and Pyramid Feature Fusion for PCB Detection. *IEEE Trans. Instrum. Meas.* 73, 5001111 (2024). <https://doi.org/10.1109/TIM.2023.3326460>

21. Chen, W., Huang, Z., Mu, Q., Sun, Y.: PCB Defect Detection Method Based on Transformer-YOLO. *IEEE Access*. 10, 129480–129489 (2022). <https://doi.org/10.1109/ACCESS.2022.3228206>
22. Yu, H., Maneetham, D., Rabgyal, T.: Performance evaluation of YOLOv12 series models for defect detection in PCB manufacturing. In: 2025 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), pp. 687–694. *IEEE* (2025). <https://doi.org/10.1109/ICIMCIS68501.2025.11326959>
23. Wen, J., Wu, Y., Huang, W., Han, K., Luo, B.: Fully cooperative domain adaptive neural network for defect classification in printed circuit boards. *Eng. Appl. Artif. Intell.* 156, 111103 (2025). <https://doi.org/10.1016/j.engappai.2025.111103>
24. Chen, S., Shao, M., Zhang, L., Bao, Z.: Two-stage structural information enhancement for source-free domain adaptation. *Mach. Vis. Appl.* 34, 121 (2023). <https://doi.org/10.1007/s00138-023-01472-5>
25. Ye, B., Lai, J., Xie, X., Zhu, J.: Prototype-guided domain adaptive one-stage object detector for defect detection. *Advanced Engineering Informatics*. 62, 102618 (2024). <https://doi.org/10.1016/j.aei.2024.102618>
26. Li, C., Pan, X., Zhu, P., Zhu, S., Liao, C., Tian, H., Qian, X., Li, X., Wang, X., Li, X.: Style Adaptation module: Enhancing detector robustness to inter-manufacturer variability in surface defect detection. *Comput. Ind.* 157–158, 104084 (2024). <https://doi.org/10.1016/j.compind.2024.104084>
27. Elmi, Z., Turan, F.A., Tuna, E., Akkaya, A., Elmi, S.: Self-adaptive deployment of synthetic-trained object detectors via test-time adaptation. *Mach. Vis. Appl.* 37, 87 (2026). <https://doi.org/10.1007/s00138-026-01849-2>
28. Pham, T.T.A., Thoi, D.K.T., Choi, H., Park, S.: Defect Detection in Printed Circuit Boards Using Semi-Supervised Learning. *Sensors*. 23, 3246 (2023). <https://doi.org/10.3390/s23063246>
29. Bilal, M., Podishetti, R., Koval, L., Gaafar, M.A., Grossmann, D., Bregulla, M.: The Effect of Annotation Quality on Wear Semantic Segmentation by CNN. *Sensors*. 24, 4777 (2024). <https://doi.org/10.3390/s24154777>
30. Baumgart, N., Lange-Hegemann, M., Mücke, M.: Investigation of the impact of synthetic training data in the industrial application of terminal strip object detection. *Mach. Vis. Appl.* 37, 13 (2026). <https://doi.org/10.1007/s00138-025-01770-0>
31. Huang, W., Wei, P.: A PCB Dataset for Defects Detection and Classification. *arXiv:1901.08204* (2019). <https://doi.org/10.48550/arXiv.1901.08204>
32. Kumar Ancha, V., Sibai, F.N., Gonuguntla, V., Vaddi, R.: Utilizing YOLO Models for Real-World Scenarios: Assessing Novel Mixed Defect Detection Dataset in PCBs. *IEEE Access*. 12, 100983–100990 (2024). <https://doi.org/10.1109/ACCESS.2024.3430329>
33. Tang, S., He, F., Huang, X., Yang, J.: Online PCB Defect Detector On A New PCB Defect Dataset. *arXiv:1902.06197* (2019). <https://doi.org/10.48550/arXiv.1902.06197>
34. Tian, Y., Ye, Q., Doermann, D.: YOLOv12: Attention-Centric Real-Time Object Detectors. In: *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)* (2025). <https://doi.org/10.52202/085713-2627>
35. Yu, H., Maneetham, D., Crisnapati, P.N., Aung, M.M.: YOLOv12-DynaFocus: A parameter-budgeted dynamic-attention detector for thin-defect-aware PCB inspection under compute and latency constraints. *IEEE Access*. 14, 130258–130300 (2026). <https://doi.org/10.1109/ACCESS.2026.3726059>
36. Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., Chen, J.: DETRs Beat YOLOs on Real-time Object Detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16965–16974 (2024). <https://doi.org/10.1109/CVPR52733.2024.01605>
37. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common Objects in Context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) *Computer Vision – ECCV 2014. Lecture Notes in Computer Science*, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
38. Kim, K., Kakani, V., Kim, H.: Automatic Pruning and Quality Assurance of Object Detection Datasets for Autonomous Driving. *Electronics (Basel)*. 14, 1882 (2025). <https://doi.org/10.3390/electronics14091882>

Supplementary Files

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

- [ESM1.pdf](#)