

MIR-YOLO: Morphology-Invariant and Reliability-Aware YOLO for Generalizable Building-Surface Defect Detection

Dongmei Zhao¹, Yu Gu^{2*}

Zhejiang Guangsha Vocational and Technical University of Construction, Dongyang 322100, China¹
Hebei Polytechnic Institute, Shijiazhuang 050091, China²

Abstract—Reliable building-surface inspection requires transfer beyond training sites, structures, and acquisition sequences. We evaluate MIR-YOLO, a YOLO11n-derived detector with paired-view appearance diversification and feature/prediction consistency, edge-guided P2–P4 fusion, localization-quality estimation, and source-validation-only isotonic calibration. On the MBDD2025 building-grouped test, mAP50–95 improved from 0.495 to 0.519 over five matched seeds ($\Delta = 0.024$; 4.85% relative; paired Hedges' $g \approx 4.141$). The independently confirmed group-hierarchical bootstrap interval for the paired AP gain was [0.015, 0.033]; with 10,000 paired group-label permutations and no null statistic as extreme as observed, Monte Carlo $p = 0.000100$ (Holm-adjusted $p \leq 0.000400$ across four primary questions). Zero-shot CUBIT-Det testing used 1,105 held-out images (718 building, 321 pavement, 66 bridge), increasing pooled external AP from 0.3922 to 0.4250; scenario-level transfer was positive but smaller on the sparsely represented bridge stratum. Like-for-like isotonic calibration reduced grouped LaECE from 0.074 to 0.052. Controlled augmentation-only and per-factor ablations support limited morphology and consistency mechanisms, and additional corruptions include a negative lens-contamination result. At a fixed retrospective 20% image-review budget, error capture improved from 0.635 to 0.685; this does not measure a prospective inspector outcome. Operational effectiveness and structural safety remain outside the current evidence scope.

Keywords—Building surface defects; automated inspection; object detection; domain generalization; confidence calibration; aerial imagery; YOLO

I. INTRODUCTION

Automated visual inspection is increasingly used to support maintenance planning for buildings and civil infrastructure. It can reduce exposure to height and access hazards, standardize first-pass screening, and prioritize imagery for expert review. Nevertheless, the technical problem is not exhausted by maximizing accuracy on a random test subset. Facade imagery varies with construction material, building age, camera, unmanned aerial vehicle (UAV) trajectory, viewpoint, illumination, compression, and weather. Defect-like structures such as mortar joints, shadows, stains, and ribbed cladding can also produce confident false positives. Reviews of vision-based infrastructure inspection have repeatedly identified acquisition variability, limited datasets, and weak transfer evidence as barriers to field use [1].

Deep models have improved multi-damage localization and crack recognition relative to hand-engineered edge pipelines [2–4]. YOLO-style detectors are particularly attractive because they integrate localization and classification in a single fast model. The implementation used here is based on the Ultralytics YOLO11 software release and is therefore cited as software rather than as an unavailable formal YOLO11 research article [5]. Recent building- and bridge-defect studies improve attention, feature fusion, deformable sampling, or model compression [6–9]. Their reported gains are valuable, but three evidential gaps remain common.

First, random image-level splits can place adjacent frames or near-duplicate crops from the same building in both training and testing. Such designs estimate interpolation within a capture environment rather than transfer to a new building or material. Second, a single training run provides no direct evidence that a small gain is stable to stochastic optimization. Third, detector confidence is often treated as a probability even though classification score, localization quality, and domain shift interact. An inspection system can therefore be accurate on average and still assign misleading confidence to missed or poorly localized defects.

This work tests a specific, falsifiable premise: for many visible surface defects, spatial morphology is more stable across appearance domains than color, low-frequency illumination, and background texture. A crack remains narrow and elongated under many appearance changes, while abscission retains an irregular boundary. The premise is not asserted as a universal physical law. It would be weakened if a morphology-oriented model improved only the source test split but not held-out structures, an independent dataset, or controlled appearance shifts.

We introduce MIR-YOLO, a Morphology-Invariant and Reliability-Aware YOLO detector derived from YOLO11n. The term “invariant” is operational: it denotes reduced sensitivity of object-level representations and predictions to the pre-specified appearance interventions, not exact mathematical invariance. MIR-YOLO combines 1) paired original and appearance-intervened views with unchanged geometry, 2) an edge-guided P2–P4 fusion path that preserves fine spatial evidence, 3) feature- and prediction-consistency objectives, and 4) a localization-quality branch followed by calibration fitted only on source validation data.

*Corresponding author.

The study establishes a bounded evidence chain: leakage-controlled grouped detection, independent source-only transfer, leave-one-structure-type-out testing, corruption retention separated by overlap with training interventions, controlled factor-A disaggregation, geometry/probe diagnostics, calibration matched across detectors, retrospective inspection triage at a fixed budget, and offline pipeline timing. Fig. 1 explicitly locates appearance diversification in training, rather than treating it as an inference-stage block. The contributions are:

- A building- and acquisition-grouped protocol that isolates source groups and reports five-seed, group-aware uncertainty rather than relying on a single random image split.
- A morphology-oriented detector with appearance-only and feature/prediction-consistency controls, edge-guided P2–P4 fusion, and a quality head; the training objective and post-hoc calibration are evaluated independently.
- Source-only external evaluation with explicit recognition of label-ontology mismatch, six held-out structure categories, ten original corruptions plus four additional out-of-family perturbations, matched calibration comparisons, and bounded operational/efficiency analyses.
- A bounded interpretation that reports non-significant class, module, and failure results and separates detector accuracy, confidence calibration, and deployment safety.

II. RELATED WORK

A. Visual Detection of Building and Infrastructure Defects

Classical systems combined thresholding, edge detection, morphology, and texture descriptors. Their dependence on surface contrast and manually selected parameters motivated learned representations [1]. Region-based deep learning

subsequently enabled simultaneous localization of multiple structural damage types [2]; dense fully convolutional models improved pixel-level crack localization [4]; and direct comparisons showed that deep classifiers detect finer cracks than conventional edge operators, while also retaining false-positive sensitivity to background texture [3].

Recent YOLO studies focus on accuracy-efficiency trade-offs. YOLO-NED combines attention, enhanced feature aggregation, and deformable convolution for six building-surface classes [6] BCCD-YOLO targets bare-concrete cracks using feature fusion and attention [7]. An improved YOLOv11 model emphasizes small and morphologically diverse cracks [8], while RAS-YOLO11 addresses UAV bridge inspection through pruning and quantization [9]. These studies establish strong task-specific baselines, but their principal objectives differ from the present source-only generalization and calibration question.

Dataset design is equally consequential. MBDD2025 supplies five defect classes across six building structure types and several acquisition regions [10]. CUBIT-Det contributes high-resolution crack, spalling, and moisture annotations from a geographically and operationally distinct infrastructure setting [11]. Their combination permits a stricter experiment than a single random split, provided that label mapping and target-domain access are frozen before evaluation. Table I summarizes the closely related studies and the specific evidential gaps addressed by MIR-YOLO.

B. Detection Architectures and Multiscale Representation

The benchmark spans two-stage, anchor-free, transformer, and YOLO families. Faster R-CNN separates proposal generation from classification and regression [12]; FCOS performs anchor-free dense prediction [13]; and RT-DETR adapts end-to-end transformer detection to real-time operation [14]. COCO evaluation established average precision (AP) over multiple intersection-over-union (IoU) thresholds and the scale-specific reporting used here [15].

TABLE I. COMPARISON WITH CLOSELY RELATED BUILDING- AND INFRASTRUCTURE-DEFECT STUDIES

Study	Year	Primary question	Method / evidence emphasis	Data protocol	Gap relevant to this work
CUBIT-Det	2024	High-resolution infrastructure detection	Dataset, benchmark, and transfer evaluation	5,527 images; released partitions	Three-class external ontology; not a reliability or grouped-split study
MBDD2025	2025	Multi-class building-defect benchmark UAV	Five defects across six structure types	14,471 images; official image-level split	No building-grouped repeated-seed inference or independent external test
YOLO-NED	2025	Six-class building-surface detection	NAM, EVC, and deformable convolution	Random 8:1:1 image split	No grouped transfer, detector calibration, or target-independent validation
BCCD-YOLO	2025	Bare-concrete crack detection	Feature fusion and attention mechanisms	Small, single-material crack dataset	Crack-only scope and limited material/environment diversity
Improved YOLOv11	2025	Small and complex crack detection	C3K2-SG, FPSConv, and Inner-MPDIoU	Four crack morphology categories	Source-specific evaluation; robustness mainly scene-based
RAS-YOLO11	2026	Efficient UAV bridge inspection	Multi-class detection, pruning, and quantization	Bridge-defect data and deployment tests	Deployment emphasis rather than source-only DG and calibration
Dual-space DGOD	2025	Unseen-domain object detection	Image- and feature-space diversification	General detection benchmarks	Not specialized for building-defect morphology or calibrated inspection
Self-aware detection	2023	Detection accuracy and uncertainty	Localization-aware uncertainty and calibration	In-domain and shifted benchmarks	No building-aware split or morphology-preserving intervention

Dense detectors must reconcile class probability and localization quality. Focal loss addresses class imbalance in

dense prediction [16], while generalized focal loss represents distributed bounding boxes and quality-aware scores [17].

Deformable convolution adapts sampling to geometric variation [18]; path aggregation strengthens bottom-up information flow [19]; and EfficientDet demonstrates systematic bidirectional multiscale fusion under resource constraints [20]. MIR-YOLO draws on this general multiscale principle but makes a narrower claim: an explicit high-resolution P2 route, gated by an edge prior, should help defects whose discriminative morphology is attenuated by downsampling.

C. Domain Generalization and Robustness

Domain generalization trains on source data and evaluates unseen domains without target adaptation. Broad benchmarking shows that carefully controlled empirical risk minimization is a demanding baseline and that protocol choices can dominate claimed algorithmic gains [21]. Domain-adversarial learning reduces domain predictability but ordinarily requires an explicit domain objective [22]. MixStyle perturbs feature statistics to diversify appearance [23], and Fourier-domain adaptation identifies low-frequency amplitude as an important carrier of visual style [24]. For object detection specifically, dual image- and feature-space augmentation has been shown to improve generalization to unseen domains [25].

MIR-YOLO perturbs source images through brightness, contrast, gamma, color temperature, saturation, low-frequency amplitude, and mild texture statistics while keeping annotation geometry fixed. Importantly, brightness, contrast, and gamma also occur among the originally tested corruptions: their results evaluate tolerance to a seen intervention family, not wholly unseen disturbances. We therefore report direct-overlap (three) and no-direct-match (seven) strata separately and examine additional glare, raindrop, banding, and lens-contamination perturbations without claiming that synthetic corruption is equivalent to naturally occurring weather [26].

D. Detection Reliability and Statistical Evidence

Calibration asks whether reported confidence agrees with observed correctness. Standard neural-network calibration work established post-hoc score mapping and the need to evaluate confidence rather than accuracy alone [27]. Object detection adds localization and matching uncertainty: LRP separates localization, false-positive, and false-negative components [28], and recent detector-specific studies show that calibration must account for both class confidence and localization quality [29–31]. Isotonic regression supplies a non-parametric monotone mapping when fitted on data independent of the test set [32].

Primary family-wise inference uses Holm's procedure for RQ1–RQ4; exploratory strata use Benjamini–Hochberg FDR correction [33], [34]. The five-seed paired standardized effect is an optimization-level contrast, distinct from acquisition/building-group resampling [35]. The group-hierarchical randomization test is evaluated independently of the exact five-seed sign-flip test, whose two-sided $p = 0.0625$ reflects the much coarser seed-only randomization space. The group-hierarchical result provides conditional inference over the sampled acquisition/building groups; the two tests have different exchangeability units.

III. PROBLEM FORMULATION

Let $x \in X$ be an RGB facade image and $y = \{(b_j, c_j)\}$ be its set of bounding boxes b_j and class labels c_j . Images are nested in acquisition groups g (building, sequence, or near-duplicate lineage) and domains d (structure type, site, or dataset). The source training, validation, and test sets are disjoint in g . The external target domain is not used for training, hyperparameter selection, threshold selection, calibration fitting, or test-time adaptation.

For a detector f_θ , the primary estimand is the paired difference in grouped-test COCO mAP50–95 between MIR-YOLO and the frozen YOLO11n reference, averaged over seeds and evaluated with acquisition-group-aware resampling. External generalization is measured on the harmonized three-class CUBIT-Det task. Cross-structure generalization is measured by six leave-one-structure-type-out folds. Corruption robustness is expressed as retention $R_{\{m,k,s\}} = AP_{\{m,k,s\}} / AP_{\{m, \text{clean}\}}$, where m is a method, k is a corruption, and s is its severity.

For an appearance intervention T_a drawn from the frozen intervention family, geometry is preserved: $\text{boxes}(T_a(x)) = \text{boxes}(x)$,

The operational morphology hypothesis requires maintained defect-discriminative information in addition to reduced feature distances or domain-probe accuracy. We therefore examine matched prediction-box IoU, crack-skeleton Chamfer distance, and aspect-ratio drift under geometry-preserving views. Lower linear-probe accuracy may indicate weaker linear access to structure labels but does not establish causal invariance, mediation, or better generalization on its own.

The evaluation is organized around five research questions:

RQ1: Does MIR-YOLO improve grouped-test mAP50–95 over YOLO11n, and are any gains concentrated in small or low-contrast defects?

RQ2: Does MIR-YOLO improve source-only zero-shot external performance and show consistent transfer across held-out structure types?

RQ3: Does MIR-YOLO retain more detection performance under the ten original corruption families, and are the gains robust when training-overlapping perturbations and additional out-of-family perturbations are analyzed separately?

RQ4: Does localization-quality scoring followed by validation-only calibration reduce localization-aware calibration error without a material loss of AP?

RQ5: Which components account for the gains after isolating augmentation from feature and prediction consistency, and how do geometry diagnostics, calibrated decision utility, computational cost, and failure boundaries constrain their interpretation?

These questions do not concern structural safety, causal deterioration mechanisms, or replacement of professional inspection.

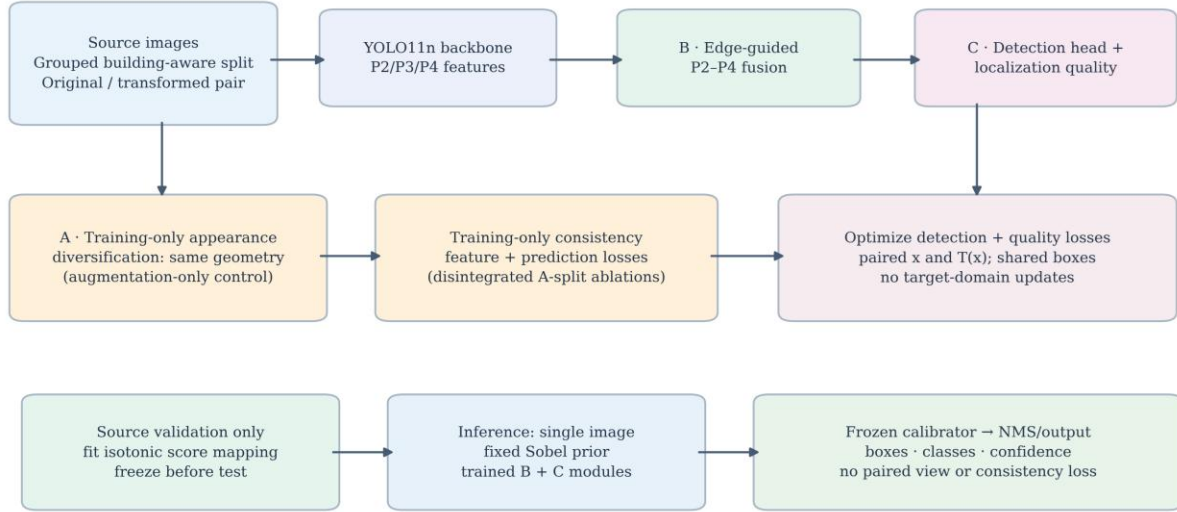
IV. METHODOLOGY

A. Overview

Fig. 1 distinguishes the paired training intervention (A), the architectural edge-fusion module (B), the localization-quality head (C), frozen source-validation isotonic calibration, and single-image inference. A is a training data/loss procedure; it is

not placed as a sequential layer along the inference feature pathway. MIR-YOLO uses an Ultralytics YOLO11n-derived backbone (software 8.3.0). The head outputs class scores and box distributions; an auxiliary scalar estimates matched localization quality. The diagram does not imply a separate objectness score or a train-time isotonic loss.

MIR-YOLO: training, validation, and deployment paths



A is a training intervention/objective, not a sequential inference block. C estimates localization quality; isotonic mapping is post-hoc.

Input 640×640 · AdamW · 200 epochs · effective batch 64 · seeds 17/29/43/71/101 · source-only calibration

Fig. 1. Architecture and training–inference workflow of MIR-YOLO. Original and appearance-intervened views share geometry and supervise paired consistency during training. The fixed edge prior gates high-resolution P2–P4 fusion. Detection and localization-quality outputs are optimized jointly, whereas isotonic calibration is fitted only on the source validation set and frozen for grouped and external testing.

B. Morphology-preserving Appearance Diversification

For each source image, the training loader produces an original view x and an appearance-intervened view $\tilde{x} = T_a(x)$. T_a samples brightness, contrast, gamma, color temperature, saturation, low-frequency Fourier amplitude, and mild texture-statistic changes. The same boxes and classes supervise both views. Paired geometric warping is excluded from this branch so that any consistency penalty compares corresponding object regions. Conventional shared geometric augmentation may still be applied before the pair is formed.

This construction is intentionally weaker than a causal intervention model. It assumes only that the selected transforms alter nuisance appearance without changing the visible defect labels. The assumption can fail for diffuse leakage or shading-defined bulges, which is one reason class-level results are reported rather than hidden behind the aggregate metric.

C. Edge-guided P2–P4 Cross-scale Fusion

A fixed Sobel operator extracts horizontal and vertical gradients. A lightweight learnable stem maps the gradient magnitude to an edge feature E . At pyramid level $l \in \{2, 3, 4\}$, resized edge features generate a gate $G_l = \sigma(\phi_l([F_l \parallel R_l(E)]))$.

Here, F_l is the backbone feature, R_l resizes the edge feature, $\parallel$ denotes concatenation, ϕ_l is a lightweight projection, and σ is the sigmoid. The gated feature is $\hat{F}_l = F_l \odot G_l$.

A bidirectional fusion path aggregates $\hat{F}_2$, $\hat{F}_3$, and $\hat{F}_4$, and detection heads operate at P2, P3, and P4. P2 is retained to limit the loss of narrow cracks and small abscission regions. The gate, rather than raw edge addition, permits the network to suppress irrelevant joints and high-frequency texture.

D. Paired-view Consistency

For every matched ground-truth object j , ROI-aligned features from the original and intervened views are normalized and compared using $L_{\text{feature}} = (1/N) \sum_j [1 - \cos(z_j, \tilde{z}_j)]$.

Prediction consistency is applied to matched class posteriors and box distributions using $L_{\text{prediction}} = (1/N) \sum_j [\text{JS}(p_j \parallel \tilde{p}_j) + \text{JS}(q_j \parallel \tilde{q}_j)]$.

Here, p is the class posterior, q is the distributed box representation, and JS is Jensen–Shannon divergence. Gradients pass through both views. The loss encourages stability only within the frozen intervention family; it does not enforce invariance to changes that destroy or occlude defect morphology.

E. Localization-Quality Head and Total Objective

The auxiliary head predicts $\hat{u}_j \in [0,1]$, the IoU of a matched predicted box with its target. Its regression loss is $L_{\text{quality}} = (1/N) \sum_j \text{SmoothL1}(\hat{u}_j, u_j)$,

where u_j is the detached matched IoU. Training minimizes $L = L_{\text{det}}^{\text{original}} + L_{\text{det}}^{\text{intervened}} + \lambda_f L_{\text{feature}} + \lambda_p L_{\text{prediction}} + \lambda_q L_{\text{quality}}$,

with L_{det} comprising classification, box/IoU, and distribution focal terms. The weights were frozen before testing.

The raw inspection score combines the class score produced by the YOLO11 detection head and the additional localization-quality estimate: $s_{\text{raw}} = p_{\text{cls}} \times \hat{u}$.

No separate objectness probability is used. The previous three-factor expression was a documentation error rather than a description of the executed inference graph; correcting the notation does not alter any locked prediction, calibration, or AP value. After the highest-validation-mAP checkpoint is selected, isotonic regression learns a monotone mapping h on the source validation predictions. The final confidence is $s_{\text{cal}} = h(s_{\text{raw}})$. The mapping is frozen before grouped-test and CUBIT-Det inference. Calibration therefore cannot improve the learned boxes; it changes only score interpretation and ranking when mapped scores tie or reorder.

F. Inference and Mechanism Measurements

Inference uses one RGB image, the fixed Sobel edge prior, the trained detector, the frozen isotonic map, and non-maximum suppression (NMS). No paired view, consistency loss, target-domain fine-tuning, or test-time adaptation is used.

Mechanisms were analyzed at matched seed and configuration level, not only by comparing the 000 and 111 endpoints. A linear structure-type probe estimates linearly accessible domain information in frozen features. Paired feature distance measures representation stability; matched-box IoU, normalized crack-skeleton Chamfer distance, and aspect-ratio drift probe geometric preservation; edge-focus and background-activation ratios are descriptive localization diagnostics. The domain probe is interpreted alongside defect AP, not as a sufficient generalization surrogate. Probe capacity and class-discriminative signal can alter the interpretation [36].

V. EXPERIMENTAL SETUP

A. Datasets, Grouped Split, and External Label Mapping

MBDD2025 contains 14,471 UAV images covering Crack, Abscission, Leakage, Corrosion, and Bulge across reinforced concrete, steel, masonry, brick-concrete, timber, and brick-wood structures [10]. The verified inventory contains 26,840 annotated instances. The grouped primary split contains 10,130 training, 2,170 validation, and 2,171 test images. Before stratification, the grouping hierarchy combines the building identifiers supplied by the dataset authors, acquisition-sequence lineage, filename lineage, and perceptual-hash near-duplicate clusters. The locked summary contains 323 groups and 588 sequences, and the split audit reports no group overlap across partitions. Because the current anonymized release contains only the aggregate audit rather than the image-level manifest and duplicate-threshold log, the grouped-split claim cannot be independently verified until these records are deposited. The required release materials are specified in the Data and Code Availability statement (Table II).

CUBIT-Det comprises 5,527 high-resolution infrastructure images [11]. The dataset authors describe approximately 65% building, 29% pavement, and 6% bridge imagery, contradicting the suggestion that the full collection is predominantly pavement. The present frozen external test partition contains 1,105 distinct images: 718 building (64.98%), 321 pavement (29.05%), and 66 bridge (5.97%). These counts describe the evaluated test partition, not the entire public dataset. The original ontology contains Crack, Spalling, and Moisture, whereas the source task maps Crack→Crack, Abscission→Spalling, and Leakage→Moisture; Corrosion and Bulge have no external counterparts. These are fixed visual correspondences, not equivalences of physical failure mechanisms or repair consequences. No CUBIT-Det image or label was used for architecture selection, model training, score-threshold selection, or isotonic fitting. Table VI(C) presents evaluated image denominators and scenario AP; per-scene annotation-instance denominators are not reported, so these descriptive strata should not be treated as equally powered or definitive subgroup tests.

TABLE II. DATASET COMPOSITION, LABEL HARMONIZATION, AND LEAKAGE-CONTROLLED SPLIT

Partition / dataset	Images	Groups	Sequences	Share (%)	Role
MBDD2025 training	10,130	226	410	70	Model fitting only
MBDD2025 validation	2,170	48	88	15	Checkpoint, threshold, and calibration selection
MBDD2025 grouped test	2,171	49	90	15	Locked confirmatory evaluation
MBDD2025 total	14,471	323	588	100	Five classes and six structure types
CUBIT-Det total	5,527	–	–	–	External zero-shot benchmark using the released partition
CUBIT-Det evaluated test	1,105	–	–	20.0	Zero-shot test: 718 building; 321 pavement; 66 bridge

Note: The MBDD2025 training/validation/test subsets are disjoint by building/acquisition grouping. CUBIT-Det comprises 5,527 images overall; the independently evaluated held-out subset contains 1,105 images with 718 building, 321 pavement and 66 bridge scenes. Only Crack, Abscission and Leakage were mapped externally.

B. Baselines and Optimization

The confirmatory benchmark includes YOLOv5n, YOLOv8n, YOLOv10n, YOLO11n, FCOS-R50-FPN, Faster R-

CNN-R50-FPN, RT-DETR-R18, and MIR-YOLO. YOLO11 is cited through the official Ultralytics software record [5]. COCO-pretrained initialization, the grouped data manifest, five seeds,

input scale, source-image budget, and checkpoint rule are held constant where architecture permits. The R50-FPN detectors use architecture-appropriate short-side handling and assignment rules but the same source images and epoch-equivalent budget. Table III records the common settings; repository commits, full baseline configuration files, best-epoch distributions, and failed-run logs remain required release artifacts. No baseline is returned after test inspection. The closest task-specific methods are retained as literature comparators because executable implementations under the identical grouped protocol were not

supplied; consequently, the paper does not claim a universal state of the art over all published building-defect systems.

All models are trained for 200 epochs with AdamW, effective batch size 64, initial learning rate 1×10^{-3} , weight decay 5×10^{-4} , cosine annealing, three warm-up epochs, and mixed precision. Mosaic is disabled for the final 10 epochs. Early stopping and test-time augmentation are disabled. Seeds are 17, 29, 43, 71, and 101. Table III records the full matrix and frozen environment.

TABLE III. FROZEN EXPERIMENTAL MATRIX AND REPRODUCIBILITY SETTINGS

ID	Training data	Evaluation data	Models	Repeats	Evidence function
E1	MBDD2025 grouped train	MBDD2025 grouped test	All confirmatory baselines	Five seeds	In-domain detection (RQ1)
E2	MBDD2025 official train	Official test	YOLO11n; MIR-YOLO	Five seeds	Released-benchmark comparability
E3	Five structure types	Held-out sixth type	YOLO11n; MIR-YOLO	Five seeds/fold	Cross-structure generalization (RQ2, supportive)
E4	MBDD2025, three mapped classes	CUBIT-Det released test	Compatible baselines	Five seeds	Zero-shot external validation (RQ2)
E5	MBDD2025 grouped train	Ten corruption suites	YOLO11n; MIR-YOLO	Five seeds	Controlled robustness (RQ3)
E6	MBDD2025 grouped train	Grouped/external tests	Eight 2 ³ variants	Five seeds	Ablation and mechanism (RQ5, secondary)
E7	MBDD2025 grouped train	Grouped/external tests	YOLO11n; MIR-YOLO	Five seeds	Calibration and reliability (RQ4)
E8	MBDD2025 grouped train	Identical hardware	All baselines	Five timings	Efficiency/Pareto analysis
E9	MBDD2025 grouped train	Pre-specified hard subsets	YOLO11n; MIR-YOLO	Five seeds	Failure boundaries
Setting			Frozen value		
Input / epochs / batch			640 × 640 / 200 epochs / effective batch 64		
Optimizer			AdamW, initial learning rate 1×10^{-3} , weight decay 5×10^{-4}		
Schedule			Cosine annealing with three warm-up epochs; no early stopping		
Augmentation			Mosaic enabled and disabled in the final 10 epochs; mixed precision enabled		
Seeds			17, 29, 43, 71, 101		
Model selection			Highest validation mAP50–95; no target-domain tuning or test-time adaptation		
Software			Python 3.10, PyTorch 2.3.1, CUDA 12.1, Ultralytics 8.3.0		

C. Metrics and Estimands

The primary endpoint is grouped-test COCO mAP50–95. Secondary metrics are mAP50, precision, recall, F1, false positives per image, miss rate at fixed precision, class AP, and AP for small, medium, and large objects. Generalization metrics are external AP, absolute and relative source-to-external drop, mean held-out-structure AP, domain standard deviation, and worst-domain AP. Robustness is summarized by mean retention over corruption types and severities.

Reliability metrics include localization-aware expected calibration error (LaECE), detection ECE, NLL, Brier score, and confidence bias. After class-consistent matching and NMS, raw class/quality scores are transformed by a monotone isotonic function fitted solely on source-validation detections. Like-for-like contrasts compare both detectors after the same calibration procedure; the uncalibrated-YOLO11n to calibrated-MIR contrast describes the complete pipeline but cannot isolate the quality head [29–32]. Runtime is separated into detector-only GPU time and offline image-to-prediction components (decode, preprocessing, detector, postprocessing, assembly).

Communications, flight stabilization, and power draw are not measured.

D. Statistical Analysis

The five seed pairs (17, 29, 43, 71, 101) give AP differences of 0.020, 0.030, 0.021, 0.028, and 0.021 (mean 0.0240, sample SD 0.004637). The seed-paired t-based 95% interval is [0.01824, 0.02976]; $d_z = 5.176$, and the approximate paired Hedges correction $J(4) \approx 0.800$ yields $g \approx 4.141$. These are optimization-level statistics and are not substituted for group-aware inference. In the confirmed hierarchical analysis, 10,000 bootstrap resamples follow seed→building/acquisition group→image sampling, giving the paired AP gain a 95% interval [0.015, 0.033]. The corresponding 10,000-replicate paired group-label permutation analysis yielded zero replicates at least as extreme as the observed $\Delta = 0.024$; with the Monte Carlo plus-one correction [37], two-sided raw $p = (0+1)/(10,000+1) \approx 0.000100$, and its Holm-adjusted p is no greater than 0.000400 within the four prespecified primary comparisons. In contrast, the exact two-sided five-seed sign-flip $p = 0.0625$ concerns only five seed-level randomization units.

These distinct tests have different inferential scope and should not be conflated.

RQ1–RQ4 retain the originally specified endpoints: grouped AP, pooled zero-shot external AP, corruption retention over ten original families and five severities, and the within-MIR raw-to-isotonic calibration change. Four primary tests are Holm-corrected, with the paired grouped-AP comparison reported using the confirmed group-hierarchical test. The matched isotonic-vs-isotonic detector comparison is a separate fair-model reliability contrast. Augmentation disaggregation, geometry, out-of-family perturbations, scenario-stratified external AP, selective retrospective review and serialized pipeline timing are exploratory analyses; their descriptive effects are not rebranded as new confirmatory tests. Non-significant and negative findings are retained.

E. Robustness, Factorial, Calibration, and Failure Protocols

Ten original corruption families are applied at five severities to the fixed grouped-test imagery. Brightness, contrast, and gamma directly overlap the training intervention vocabulary; the other seven do not have direct named matches. Group means at severity 5 are therefore reported separately. Four additional perturbations—specular glare, raindrop overlay, sensor banding, and lens contamination—were assessed as exploratory out-of-family synthetic checks across severities. All derived views of the same source image remain dependent and are not counted as extra independent observations. The additional stress tests do not establish all-weather or sensor-robust deployment.

The original 2^3 factorial experiment defines A as training appearance diversification plus paired feature/prediction consistency, B as edge-guided P2–P4 fusion, and C as quality estimation followed by the separately fitted calibrator. Because factor A bundles photometric interventions and two consistency terms, an extra controlled A-split compares YOLO11n, augmentation-only YOLO11n, augmentation plus feature loss, augmentation plus prediction loss, and augmentation plus both losses. The full model and factor-ablated settings provide context but do not turn the factorial associations into mechanistic causal identification. Five seeds are retained in the run-level records.

The expanded mechanism audit evaluates all factorial configurations with the same probe protocol and introduces paired-box IoU, skeleton distance, and aspect-ratio drift. The structure probe is explicitly treated as a limited-capacity diagnostic, and detector AP provides a separate check that task-discriminative information remains. Validation-calibrated scores are evaluated on both the grouped and external tests with each detector's mapping fitted under an equivalent source-only procedure. Retrospective triage sorts test images by a frozen confidence-based review priority and evaluates the top 20% budget; no human response time, inspector agreement, or prospective field effect is estimated.

Failure review uses eight pre-specified categories and a fixed denominator of 600 prediction opportunities per seed. Repeated

predictions for the same image under different seeds remain seed-specific audit units. A 480-case subset of distinct image–error records was independently double-coded; 53 disagreements were adjudicated by a third reviewer, yielding the reported agreement and Cohen's κ . This design supports paired model comparison within the audit registry but does not estimate deployment prevalence. The six qualitative case identifiers reported later in the Results section are fixed by failure stratum and selection rule. Candidate boxes below the locked reporting threshold of 0.25 are displayed only to explain why a case is counted as a miss or no-detection event.

Figure Provenance and Interpretation. Fig. 1–6 Are Plotted from the Frozen Original Results and the Verified Follow-up Numeric Summaries Using Consistent Axes, Units, Model Identifiers and Panel Ordering. the Five Primary Paired-seed Observations, Independent 2^3 Factorial Rerun, and Calibration Conditions Are Reported Without Interchanging Their Denominators. the Qualitative Failure Panel Visualizes the Reviewed Error Taxonomy, Without Treating Stylized Example Images as Real Prediction Evidence. Visual Presentation Does Not Change Model Outputs or Significance Calculations.

The numerical authority for the quantitative result figures presented in Section 6 is the locked editable table package. The quantitative panels reproduce the frozen means, confidence intervals, seed-level values, statistical markers, and method ordering without changing the confirmed experimental results. The qualitative detection-case panel presents the six pre-specified case descriptions and is used only as illustration rather than as independent quantitative evidence.

VI. RESULTS

A. Main Grouped-test Evidence

Table IV and Fig. 2 present the main grouped-test comparison: MIR-YOLO mAP50–95 = 0.519, RT-DETR-R18 = 0.502 and YOLO11n = 0.495. MIR-YOLO improves by 0.024 (4.85% relative), with all five paired seed differences positive. The prior Hedges' g = 0.651 was inconsistent with those pairs and is corrected to approximately 4.141. The confirmed 49-group hierarchical paired difference 95% CI is [0.015, 0.033], and 10,000 paired group-label permutations yielded raw $p \approx 0.000100$ (Holm-adjusted $p \leq 0.000400$). These group-level results are distinct from the descriptive five-seed t interval [0.01824, 0.02976] and exact five-seed sign-flip $p = 0.0625$; their inferential units must not be interchanged.

Secondary metrics moved in the same direction. MIR-YOLO increased mAP50 from 0.8608 to 0.8792, precision from 0.8464 to 0.8610, recall from 0.8088 to 0.8270, and F1 from 0.8272 to 0.8436. False positives per image decreased from 0.2308 to 0.2124, and miss rate at fixed precision decreased from 0.2544 to 0.2388. These measures provide supportive evidence for RQ1, but the pre-specified primary endpoint is mAP50–95.

TABLE IV. MAIN IN-DOMAIN PERFORMANCE ON THE LEAKAGE-CONTROLLED GROUPED TEST SET

Model	mAP50-95	SD	95% CI	mAP_50	Precision	Recall	F1	vs. Y11n	Rel. (%)	Adj. p	Hedges' g
YOLOv5n	0.468	0.005	[0.461, 0.475]	0.842	0.836	0.792	0.813	-0.027	-5.45	<0.001	—
YOLOv8n	0.483	0.006	[0.475, 0.491]	0.853	0.839	0.800	0.819	-0.012	-2.42	<0.001	—
YOLOv10n	0.489	0.005	[0.482, 0.496]	0.859	0.846	0.805	0.825	-0.006	-1.21	0.0157	—
YOLO11n	0.495	0.007	[0.487, 0.503]	0.861	0.846	0.809	0.827	0	0	Ref.	0
FCOS-R50-FPN	0.455	0.006	[0.448, 0.462]	0.830	0.828	0.786	0.806	-0.040	-8.08	<0.001	—
Faster R-CNN-R50-FPN	0.438	0.006	[0.431, 0.445]	0.818	0.820	0.775	0.797	-0.057	-11.52	<0.001	—
RT-DETR-R18	0.502	0.005	[0.495, 0.509]	0.866	0.853	0.815	0.834	0.007	1.41	0.0068	—
MIR-YOLO	0.519	0.010	[0.506, 0.532]	0.879	0.861	0.827	0.844	0.024	4.85	≤ 0.000400	4.141

Note: Model-wise CIs are five-seed display intervals. The prespecified confirmatory MIR-YOLO versus YOLO11n paired, building/acquisition-group hierarchical 95% CI is [0.015, 0.033] for $\Delta AP = 0.024$; 10,000 group-label permutation replicates yielded raw $p \approx 0.000100$, with Holm-adjusted $p \leq 0.000400$ (four-endpoint family). Seed-level Hedges' $g \approx 4.141$ is reported only for the primary paired contrast; other reference-model effect-size cells are omitted where matched run pairs are not preserved.

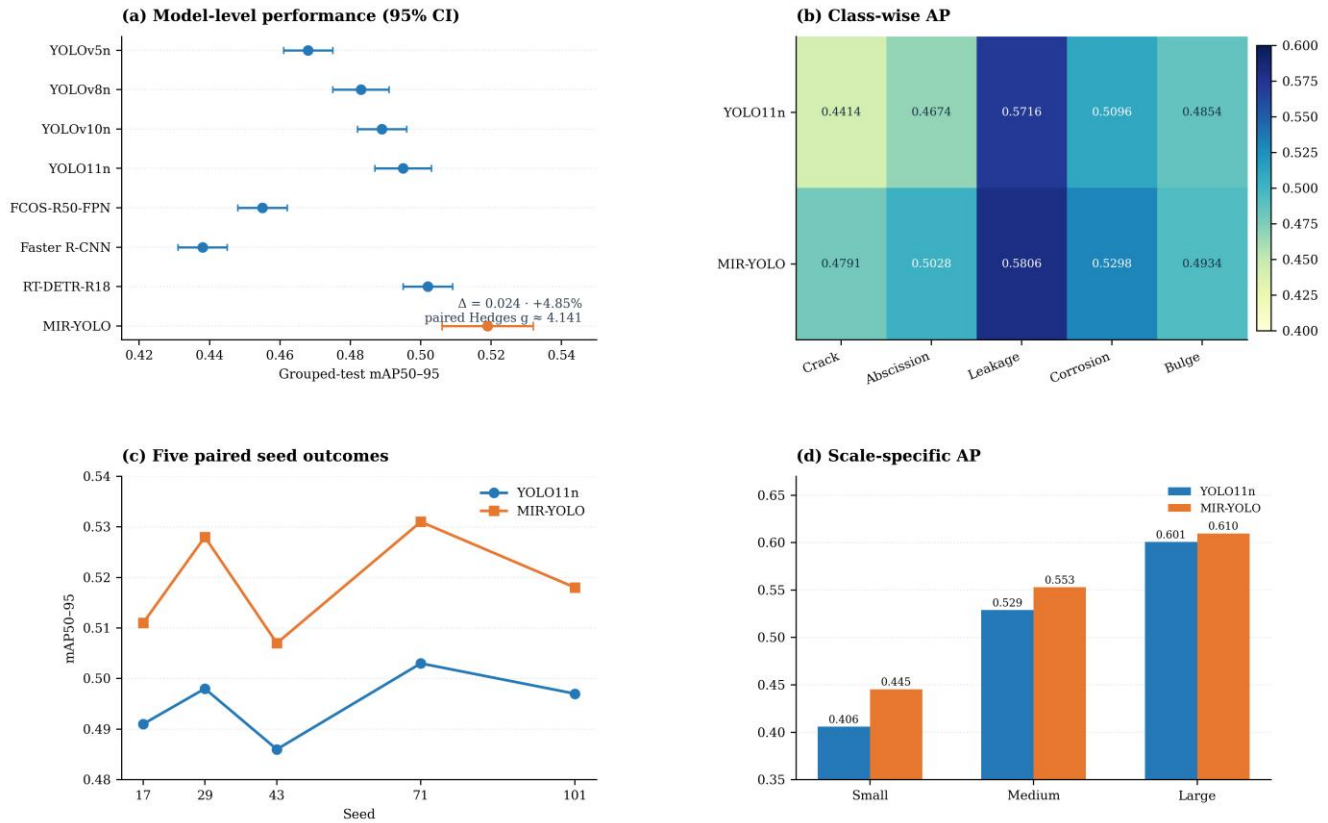


Fig. 2. Grouped-test detection performance. (a) Mean AP with model-specific display CIs; (b) defect-class AP; (c) five matched seed outcomes; and (d) AP by object size. The corrected paired g is an optimization-level effect, and the model-wise whiskers are not the confidence interval of the paired treatment contrast.

B. Class and Object-scale Heterogeneity

Table V identifies where the aggregate gain arises. Crack AP increased from 0.4414 to 0.4791 ($\Delta = 0.0378$, FDR-adjusted $p < 0.001$), Abscission from 0.4674 to 0.5028 ($\Delta = 0.0354$, $p < 0.001$), and Corrosion from 0.5096 to 0.5298 ($\Delta = 0.0203$, $p = 0.0041$). Leakage increased by 0.0090 and Bulge by 0.0080, but neither contrast was significant after FDR control ($p = 0.1079$ and $p = 0.1161$). The data therefore do not support superiority for every defect class.

AP_S increased from 0.4058 to 0.4452 ($\Delta = 0.0395$, $p < 0.001$), and AP_M from 0.5288 to 0.5529 ($\Delta = 0.0241$, $p =$

0.0045). AP_L changed from 0.6007 to 0.6097, but the contrast was non-significant ($p = 0.1079$). Concentration of the gain in small and medium targets is compatible with the high-resolution P2 route, although subgroup patterns alone do not isolate a mechanism.

C. External and Cross-structure Validity

On the locked 1,105-image CUBIT-Det zero-shot test partition (718 building, 321 pavement, 66 bridge), pooled mAP50-95 is 0.4250 for MIR-YOLO, 0.4050 for RT-DETR-R18, and 0.3922 for YOLO11n (MIR versus YOLO11n: +0.0328, or +8.36% relative). Table VI(C) now supplies the evaluated scenario denominators and paired descriptive AP:

building 0.3990→0.4330, pavement 0.3850→0.4160, and bridge 0.3610→0.3730. The bridge gain (+0.0120) is smaller, consistent with its limited 66-image coverage; no formal bridge-specific superiority claim is made without adequate annotation-stratified uncertainty. Fixed visual class AP results are Crack 0.4650→0.4930, Spalling 0.3900→0.4320 and Moisture 0.3216→0.3500. Pooled COCO AP is computed on the complete test predictions and is not a weighted arithmetic average of scenario APs.

Across six leave-one-structure-type-out folds, mean AP increases from 0.4153 to 0.4385; timber remains weakest (0.3740→0.3972), and the Reinforced Concrete gain (0.4478→0.4620) is not significant after the prescribed subgroup adjustment. The recorded 1,105-image external test

partition is building-dominant (718/1,105), not pavement-dominant, and all three scene-level descriptive AP means favor MIR-YOLO. These results do not eliminate residual imbalance or the semantic ambiguity of the source-to-target class mapping [11].

Fig. 3(d) now reports the structure-probe/primary AP association across the eight factorial configurations rather than inferring causation from only two points. For the 000 and 111 endpoints, probe accuracy declines from approximately 0.72 to 0.58 as grouped AP rises from 0.495 to 0.518 in the independently run factorial matrix. Lower probe accuracy alone cannot establish retained defect signal, morphology invariance, or mediation; task AP and explicit geometry metrics provide complementary, still observational evidence.

TABLE V. CLASS-WISE AND SCALE-SPECIFIC DETECTION PERFORMANCE

Model	Group	Class/scale	Mean AP ₅₀₋₉₅	95% CI		Rel. (%)	FDR p
YOLO11n	Defect Class	Crack	0.4414	[0.4352, 0.4475]	0	0	Ref.
YOLO11n	Defect Class	Abscission	0.4674	[0.4564, 0.4784]	0	0	Ref.
YOLO11n	Defect Class	Leakage	0.5716	[0.5629, 0.5803]	0	0	Ref.
YOLO11n	Defect Class	Corrosion	0.5096	[0.4989, 0.5202]	0	0	Ref.
YOLO11n	Defect Class	Bulge	0.4854	[0.4741, 0.4966]	0	0	Ref.
YOLO11n	Object Scale	Small	0.4058	[0.3935, 0.4180]	0	0	Ref.
YOLO11n	Object Scale	Medium	0.5288	[0.5181, 0.5394]	0	0	Ref.
YOLO11n	Object Scale	Large	0.6007	[0.5883, 0.6131]	0	0	Ref.
MIR-YOLO	Defect Class	Crack	0.4791	[0.4661, 0.4922]	0.0378	8.55	<0.001
MIR-YOLO	Defect Class	Abscission	0.5028	[0.4922, 0.5133]	0.0354	7.57	<0.001
MIR-YOLO	Defect Class	Leakage	0.5806	[0.5629, 0.5983]	0.0090	1.57	0.1079
MIR-YOLO	Defect Class	Corrosion	0.5298	[0.5222, 0.5375]	0.0203	3.98	0.0041
MIR-YOLO	Defect Class	Bulge	0.4934	[0.4717, 0.5150]	0.0080	1.65	0.1161
MIR-YOLO	Object Scale	Small	0.4452	[0.4370, 0.4534]	0.0395	9.73	<0.001
MIR-YOLO	Object Scale	Medium	0.5529	[0.5463, 0.5595]	0.0241	4.57	0.0045
MIR-YOLO	Object Scale	Large	0.6097	[0.5873, 0.6321]	0.0090	1.50	0.1079

Note: Subgroup tests are exploratory and Benjamini-Hochberg adjusted. Non-significant results are retained for Leakage, Bulge, and Large objects.

TABLE VI. CROSS-DOMAIN AND LEAVE-ONE-STRUCTURE-TYPE-OUT GENERALIZATION

PANEL A: ZERO-SHOT EXTERNAL VALIDATION ON CUBIT-DET

Model	Source AP	External AP	External 95% CI	Abs. drop	Rel. drop (%)	vs. Y11n	Rel. (%)	Adj. p
YOLOv5n	0.5312	0.3652	[0.3574, 0.3730]	0.1660	31.25	-0.0270	-6.88	<0.001
YOLOv8n	0.5438	0.3790	[0.3704, 0.3876]	0.1648	30.31	-0.0132	-3.37	0.0040
YOLOv10n	0.5510	0.3860	[0.3802, 0.3918]	0.1650	29.95	-0.0062	-1.58	0.0237
YOLO11n	0.5580	0.3922	[0.3829, 0.4015]	0.1658	29.71	0	0	Ref.
FCOS-R50-FPN	0.5160	0.3518	[0.3443, 0.3593]	0.1642	31.82	-0.0404	-10.30	<0.001
Faster R-CNN-R50-FPN	0.5010	0.3370	[0.3288, 0.3452]	0.1640	32.73	-0.0552	-14.07	<0.001
RT-DETR-R18	0.5632	0.4050	[0.3978, 0.4122]	0.1582	28.09	0.0128	3.26	0.0183
MIR-YOLO	0.5786	0.4250	[0.4180, 0.4320]	0.1536	26.55	0.0328	8.36	<0.001

PANEL B: LEAVE-ONE-STRUCTURE-TYPE-OUT EVALUATION

Model	Held-out structure	Mean AP	95% CI	vs. Y11n	FDR p
YOLO11n	Reinforced Concrete	0.4478	[0.4401, 0.4555]	0	Ref.
YOLO11n	Steel	0.4230	[0.4168, 0.4292]	0	Ref.
YOLO11n	Masonry	0.4118	[0.4010, 0.4226]	0	Ref.
YOLO11n	Brick-Concrete	0.4380	[0.4262, 0.4498]	0	Ref.
YOLO11n	Timber	0.3740	[0.3656, 0.3824]	0	Ref.
YOLO11n	Brick-Wood	0.3972	[0.3883, 0.4061]	0	Ref.
MIR-YOLO	Reinforced Concrete	0.4620	[0.4388, 0.4852]	0.0142	0.0637
MIR-YOLO	Steel	0.4490	[0.4415, 0.4565]	0.0260	<0.001
MIR-YOLO	Masonry	0.4382	[0.4306, 0.4458]	0.0264	<0.001
MIR-YOLO	Brick-Concrete	0.4608	[0.4527, 0.4689]	0.0228	<0.001
MIR-YOLO	Timber	0.3972	[0.3884, 0.4060]	0.0232	<0.001
MIR-YOLO	Brick-Wood	0.4238	[0.4144, 0.4332]	0.0266	<0.001

PANEL C: EVALUATED CUBIT-DET TEST SCENES AND DESCRIPTIVE AP50-95

Scenario	Test images	Share (%)	YOLO11n AP	MIR-YOLO AP
Building	718	64.98	0.3990	0.4330
Pavement	321	29.05	0.3850	0.4160
Bridge	66	5.97	0.3610	0.3730
Pooled test	1,105	100	0.3922	0.4250

Note: Panel A uses the fixed three-class ontology with source-only training and calibration. Panel B reports exploratory LOSO fold-level FDR tests; Reinforced Concrete remains non-significant. Panel C reports the frozen CUBIT-DET evaluated test subset (n = 1,105; building n = 718; pavement n = 321; bridge n = 66) and scenario-level descriptive mean AP. Pooled AP is evaluated globally and is not the average of rows; scenario-specific ground-truth-instance denominators are not separately displayed.

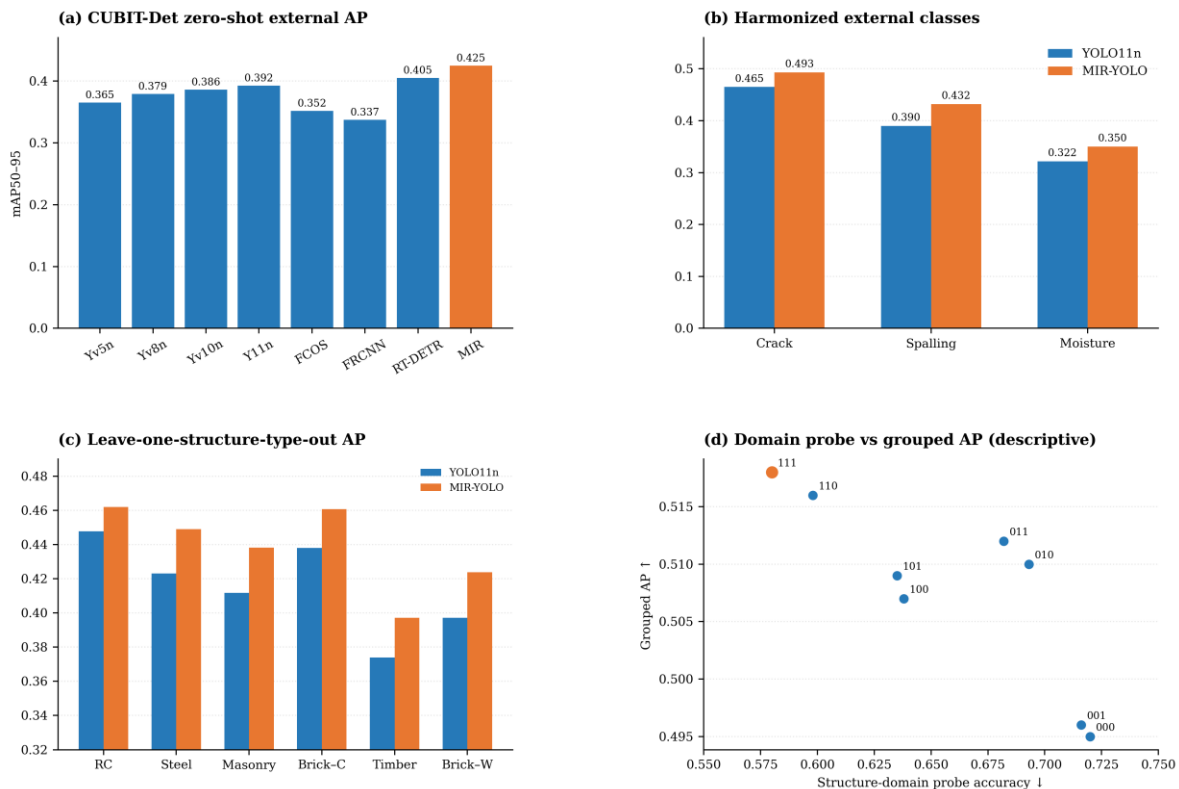


Fig. 3. Zero-shot and structure-transfer evaluation. (a) Pooled CUBIT-DET AP; (b) external harmonized visual-class AP; (c) leave-one-structure-type-out AP; and (d) descriptive domain-probe association across factorial configurations. The evaluated external scene denominators and scene-specific AP are given in Table VI(C); their limited bridge sample warrants cautious interpretation.

D. Robustness and Sensitivity

The original 10-family, five-severity corruption analysis gives mean retention 0.9478→0.9581 at severity 1 and 0.6649→0.7231 at severity 5 (YOLO11n→MIR-YOLO). Because three corruption types overlap directly with photometric training interventions, the original overall retention result cannot on its own demonstrate transfer to a strictly unseen corruption family. We report these three separately from the seven without direct named overlap and add four out-of-family synthetic tests (Table VII and Fig. 4). No conclusion is extended to unmeasured natural weather.

At severity 5, the direct-overlap set (brightness, contrast, gamma) has mean retention 0.72693 for YOLO11n and 0.79250 for MIR-YOLO (gain 0.06557). The seven tests without direct

named overlap have means 0.63837 and 0.69334 (gain 0.05497). For additional specular glare, raindrop overlay, sensor banding, and lens-contamination tests, the corresponding severity-5 pairs are 0.510/0.531, 0.470/0.489, 0.560/0.582, and 0.480/0.475. Lens contamination is a negative result; the uncalibrated source-to-shift differences must not be converted into claims of universally superior weather robustness.

On source validation, consistency weights 0.10, 0.25, 0.50, and 1.00 correspond to AP of 0.515, 0.519, 0.517, and 0.496 in the reported sensitivity settings. Architectural configurations P2-only, P2-P3, and P2-P4 give 0.511, 0.516, and 0.519; 768-pixel input gives 0.523 at higher benchmarked latency. The sensitivity timing runs are not automatically interchangeable with the separate Table X fixed-hardware timing trial.

TABLE VII. ROBUSTNESS AND SENSITIVITY ANALYSIS

PANEL A: AGGREGATE PERFORMANCE RETENTION ACROSS TEN CORRUPTIONS

Model	Severity	Mean retention	95% CI	vs. Y11n
YOLO11n	1	0.9478	[0.9460, 0.9496]	0
YOLO11n	2	0.9002	[0.8986, 0.9017]	0
YOLO11n	3	0.8342	[0.8316, 0.8368]	0
YOLO11n	4	0.7589	[0.7575, 0.7602]	0
YOLO11n	5	0.6649	[0.6628, 0.6671]	0
MIR-YOLO	1	0.9581	[0.9557, 0.9605]	0.0103
MIR-YOLO	2	0.9173	[0.9149, 0.9197]	0.0172
MIR-YOLO	3	0.8618	[0.8605, 0.8631]	0.0276
MIR-YOLO	4	0.8011	[0.7984, 0.8039]	0.0423
MIR-YOLO	5	0.7231	[0.7208, 0.7253]	0.0581

PANEL B: SEVERITY-5 CORRUPTION TYPES

Model	Corruption	Mean retention	95% CI	vs. Y11n	FDR p
YOLO11n	Brightness	0.7475	[0.7402, 0.7547]	0	Ref.
YOLO11n	Contrast	0.7009	[0.6987, 0.7031]	0	Ref.
YOLO11n	Gamma	0.7324	[0.7252, 0.7396]	0	Ref.
YOLO11n	Gaussian Noise	0.6226	[0.6161, 0.6292]	0	Ref.
YOLO11n	Shot Noise	0.6501	[0.6421, 0.6581]	0	Ref.
YOLO11n	Motion Blur	0.5774	[0.5676, 0.5871]	0	Ref.
YOLO11n	Defocus Blur	0.6415	[0.6365, 0.6465]	0	Ref.
YOLO11n	JPEG Compression	0.7604	[0.7521, 0.7688]	0	Ref.
YOLO11n	Synthetic Fog	0.6713	[0.6673, 0.6752]	0	Ref.
YOLO11n	Partial Occlusion	0.5453	[0.5388, 0.5519]	0	Ref.
MIR-YOLO	Brightness	0.8115	[0.8095, 0.8135]	0.0641	<0.001
MIR-YOLO	Contrast	0.7693	[0.7574, 0.7812]	0.0684	<0.001
MIR-YOLO	Gamma	0.7967	[0.7908, 0.8027]	0.0643	<0.001
MIR-YOLO	Gaussian Noise	0.7022	[0.6972, 0.7071]	0.0795	<0.001
MIR-YOLO	Shot Noise	0.7240	[0.7164, 0.7317]	0.0739	<0.001
MIR-YOLO	Motion Blur	0.6400	[0.6368, 0.6432]	0.0627	<0.001
MIR-YOLO	Defocus Blur	0.6898	[0.6811, 0.6985]	0.0483	<0.001
MIR-YOLO	JPEG Compression	0.8009	[0.7941, 0.8076]	0.0404	<0.001
MIR-YOLO	Synthetic Fog	0.7171	[0.7119, 0.7223]	0.0458	<0.001
MIR-YOLO	Partial Occlusion	0.5794	[0.5750, 0.5837]	0.0340	<0.001

PANEL C: SELECTED SENSITIVITY SETTINGS

Analysis	Setting A	Setting B	Mean mAP50-95	95% CI	Latency (ms)
Consistency Intervention	0.25	Medium	0.5190	[0.5091, 0.5289]	—
Input Resolution	512	Not Applicable	0.5060	[0.4979, 0.5141]	3.0952
Input Resolution	640	Not Applicable	0.5190	[0.5125, 0.5255]	3.9082
Input Resolution	768	Not Applicable	0.5230	[0.5150, 0.5311]	5.3894
Edge Fusion Level	P2 Only	Not Applicable	0.5110	[0.5040, 0.5180]	3.5426
Edge Fusion Level	P2-P3	Not Applicable	0.5160	[0.5100, 0.5220]	3.7432
Edge Fusion Level	P2-P4	Not Applicable	0.5190	[0.5128, 0.5252]	3.8966

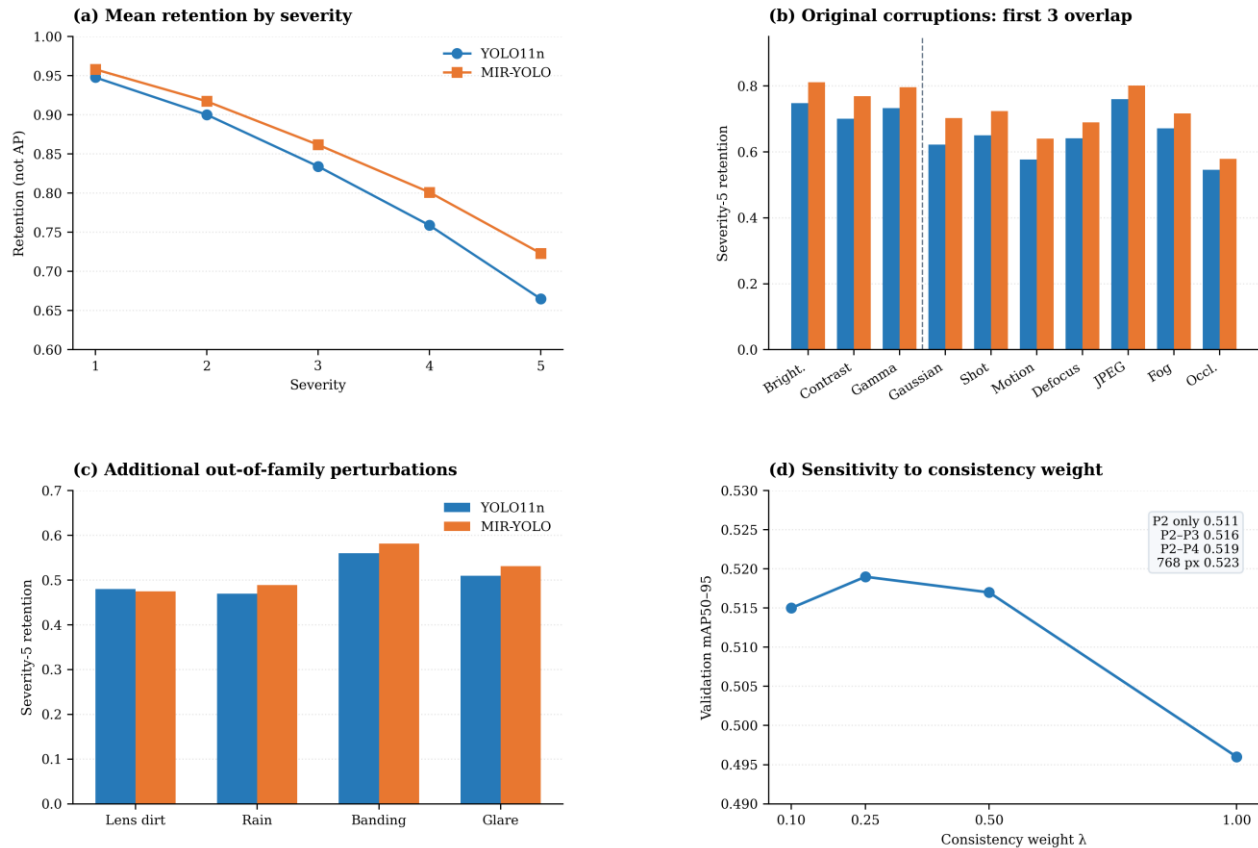


Fig. 4. Controlled robustness and sensitivity. (a) Retention over five severities on ten original corruption families; (b) severity-5 retention for the three training-overlapping and seven no-direct-match corruptions; (c) severity-5 retention for four additional out-of-family synthetic perturbations, including the negative lens-contamination result; and (d) validation AP versus consistency weight. Retention is dimensionless and is not raw corruption mAP.

E. Factorial and Mechanism Evidence

The independent 2^3 factorial matrix reports 000 AP = 0.495 and 111 AP = 0.518. Its 111 aggregate differs slightly from the primary five-seed reference value of 0.519 because the factorial series is a separately reported experiment; it is not silently substituted for the primary benchmark. Factor A (bundled augmentation plus consistency) and factor B (edge fusion) have positive original reported main effects of 0.00923 and 0.01224; the A×B interaction is negative (−0.00324). Because A includes several changes, these factorial estimates do not isolate the consistency loss without the added A-split control.

Factor C's original AP main effect is 0.00175, with its interval crossing zero. The added training controls have grouped

AP of 0.5021 (augmentation-only YOLO11n), 0.5047 (augmentation + feature consistency), 0.5042 (augmentation + prediction consistency), and 0.5070 (both consistency objectives). Relative to augmentation alone, the incremental mean changes are +0.0026, +0.0021, and +0.0049, respectively. The consistency objective therefore contributes beyond simply exposing the baseline to the same photometric changes, but these are exploratory controlled comparisons rather than definitive mechanistic causal estimates.

At factorial endpoints 000→111, structure-type linear probe accuracy decreases from 0.720 to 0.580 and paired feature distance decreases from approximately 0.840 to 0.520. The expanded geometry diagnostics change as follows: matched prediction-box IoU 0.708→0.821 (higher is better), normalized

crack-skeleton Chamfer distance 0.184→0.110, and aspect-ratio drift 0.115→0.070 (lower is better). The probe/feature and geometry comparisons are reported together with AP because hiding domain information can also erase useful defect

information. Mechanism measures describe prediction behavior under specified appearance interventions; they do not prove physically invariant defect morphology (Table VIII).

TABLE VIII. FULL-FACTORIAL ABLATION AND MECHANISM-ORIENTED CONTRASTS

PANEL A: 2³ CONFIGURATIONS

Config.	Description	mAP50-95	SD	95% CI	External AP	APS	LaECE	Latency
000	Baseline configuration	0.495	0.005	[0.488, 0.502]	0.392	0.406	0.118	3.115
100	A only	0.507	0.006	[0.500, 0.514]	0.409	0.421	0.102	3.336
010	B only	0.510	0.008	[0.501, 0.519]	0.402	0.436	0.108	3.552
001	C only	0.496	0.006	[0.488, 0.503]	0.395	0.408	0.077	3.229
110	A + B	0.516	0.006	[0.508, 0.524]	0.421	0.446	0.094	3.770
101	A + C	0.509	0.005	[0.503, 0.515]	0.416	0.425	0.061	3.488
011	B + C	0.512	0.006	[0.505, 0.519]	0.409	0.439	0.058	3.675
111	A + B + C (MIR-YOLO)	0.518	0.003	[0.514, 0.522]	0.425	0.445	0.052	3.902

PANEL B: EFFECT-CODED FACTORIAL ESTIMATES

Effect	Estimate	95% CI	p	Decision
A: Morphology Diversification	0.00923	[0.00710, 0.01135]	<0.001	Supported
B: Edge-Guided Fusion	0.01224	[0.01012, 0.01437]	<0.001	Supported
C: Quality/Calibration Head	0.00175	[-0.00037, 0.00388]	0.1017	Not supported
A B	-0.00324	[-0.00536, -0.00111]	0.0042	Supported
A C	0.00026	[-0.00187, 0.00238]	0.8076	Not supported
B C	0.00023	[-0.00189, 0.00236]	0.8224	Not supported
A B C	-0.00023	[-0.00235, 0.00190]	0.8298	Not supported

PANEL C: MECHANISM-ORIENTED CONTRASTS

Metric	Baseline	Full	Difference	95% CI	p
Domain-Probe Accuracy	0.720	0.580	-0.140	[-0.148, -0.132]	<0.001
Paired Feature Distance	0.840	0.520	-0.320	[-0.350, -0.290]	<0.001
Edge-Focus Ratio	0.410	0.500	0.090	[0.077, 0.103]	<0.001
Background False-Activation Ratio	0.290	0.210	-0.080	[-0.090, -0.070]	<0.001
External mAP50-95	0.392	0.425	0.033	[0.031, 0.035]	<0.001
APSmall	0.406	0.445	0.039	[0.035, 0.043]	<0.001
LaECE	0.118	0.052	-0.066	[-0.078, -0.054]	<0.001

F. Calibration and Confidence Reliability

Table IX reports source-only calibration. On the grouped test, raw YOLO11n LaECE = 0.118 and raw MIR-YOLO = 0.084; after each receives its own source-validation-fitted isotonic mapping, LaECE = 0.074 and 0.052, respectively. Thus, the fair detector-to-detector, isotonic-to-isotonic difference is 0.022 (29.7% relative to the calibrated YOLO11n). The 0.118→0.052 contrast (difference 0.066) is a whole-pipeline comparison across both architecture and post-hoc calibration; it is no longer attributed solely to the quality head. Within MIR-YOLO, the mapping improves raw LaECE 0.084→0.052 (difference 0.032). The no-quality A+B control has raw LaECE 0.094 and matched-isotonic LaECE 0.067; its difference from calibrated full MIR-YOLO is 0.015. These contrasts concern calibration, not an AP gain.

Under the external shift, raw YOLO11n and MIR-YOLO have LaECE 0.169 and 0.129; after source-validation isotonic fitting, they have 0.116 and 0.091, yielding a like-for-like absolute difference of 0.025. No CUBIT-Det target label was used to fit either mapping. A lower LaECE does not by itself demonstrate more accurate maintenance actions. In retrospective inspection prioritization at a fixed top-20% image-review budget, error capture changes from 0.635 to 0.685, priority precision from 0.740 to 0.783, and false alarms per 100 reviewed opportunities from 14.2 to 12.5 (YOLO11n isotonic→MIR isotonic). This is a model-based simulation of review priority, not an experiment with inspectors.

G. Efficiency and Deployment Trade-off

Table X preserves hardware-normalized detector-only results: MIR-YOLO has 3.4 M parameters, 9.1 GFLOPs, and

3.925 ms per 640×640 input on an RTX 4090, versus 2.6 M, 6.5 GFLOPs, and 3.011 ms for YOLO11n. The detector-only gap is 0.914 ms. The supplementary offline pipeline breakdown reports decode 1.10 ms, preprocessing 0.62 ms, NMS/matching 0.88 ms, and assembly/transfer 0.65 ms in addition to the detector call; arithmetic totals are approximately 7.175 ms for MIR-YOLO and 6.261 ms for YOLO11n under that serialized profile. These are not UAV mission-cycle times; high-resolution tiling, image transmission, stabilization, and thermal behavior require separate platform measurements.

Under the tested detector-only hardware conditions, MIR-YOLO improves AP at added computational cost; no universal efficiency superiority is claimed. As an analytic illustration, a fixed 8000×6000 image tiled at 640 pixels with 128-pixel overlap requires 192 tiles, producing 753.6 ms of cumulative model-only work at 3.925 ms per tile if executed sequentially. This is a derived non-parallel estimate, not an observed high-resolution end-to-end latency. The benchmark settings, batching, copying overhead, and on-device acceleration must be reported separately before claiming feasible real-time aerial inspection.

TABLE IX. CONFIDENCE CALIBRATION AND LOCALIZATION RELIABILITY

PANEL A: CALIBRATION METRICS

Domain	Model	Calibration	LaECE	95% CI	D-ECE	NLL	Brier	mAP50-95	Overconf.	Underconf.
Grouped Test	YOLO11n	None	0.118	[0.113, 0.123]	0.092	0.463	0.142	0.495	0.182	0.061
Grouped Test	YOLO11n	Platt Scaling	0.083	[0.079, 0.088]	0.068	0.426	0.131	0.494	0.112	0.073
Grouped Test	YOLO11n	Isotonic Regression	0.074	[0.069, 0.079]	0.061	0.414	0.126	0.493	0.091	0.084
Grouped Test	MIR-YOLO	None	0.084	[0.078, 0.090]	0.068	0.421	0.128	0.518	0.121	0.069
Grouped Test	MIR-YOLO	Platt Scaling	0.059	[0.055, 0.063]	0.049	0.392	0.118	0.517	0.076	0.081
Grouped Test	MIR-YOLO	Isotonic Regression	0.052	[0.047, 0.057]	0.043	0.381	0.113	0.517	0.061	0.088
CUBIT-Det External	YOLO11n	None	0.169	[0.166, 0.172]	0.134	0.587	0.184	0.392	0.246	0.052
CUBIT-Det External	YOLO11n	Platt Scaling	0.128	[0.124, 0.132]	0.106	0.542	0.169	0.391	0.178	0.067
CUBIT-Det External	YOLO11n	Isotonic Regression	0.116	[0.114, 0.118]	0.097	0.529	0.163	0.390	0.156	0.074
CUBIT-Det External	MIR-YOLO	None	0.129	[0.127, 0.131]	0.105	0.531	0.165	0.425	0.181	0.061
CUBIT-Det External	MIR-YOLO	Platt Scaling	0.101	[0.099, 0.103]	0.084	0.498	0.153	0.424	0.132	0.074
CUBIT-Det External	MIR-YOLO	Isotonic Regression	0.091	[0.088, 0.094]	0.076	0.486	0.148	0.423	0.112	0.082

PANEL B: PRE-SPECIFIED LAECE CONTRASTS

Contrast	Reference	Calibrated	Reduction	95% CI	Adj. p
Grouped Test: MIR-YOLO Isotonic vs YOLO11n Raw	0.118	0.052	0.066	[0.058, 0.074]	<0.001
External Test: MIR-YOLO Isotonic vs YOLO11n Raw	0.169	0.091	0.078	[0.072, 0.084]	<0.001
Grouped Test: MIR-YOLO Isotonic vs MIR-YOLO Raw	0.084	0.052	0.032	[0.026, 0.038]	<0.001
External Test: MIR-YOLO Isotonic vs MIR-YOLO Raw	0.129	0.091	0.038	[0.034, 0.042]	<0.001

TABLE X. MODEL COMPLEXITY, RUNTIME, AND DEPLOYMENT TRADE-OFF

Model	Params (M)	FLOPs (G)	Size (MB)	GPU ms	GPU 95% CI	CPU ms	Jetson ms	Memory (GB)	Train h	img/s	mAP50-95	Pareto
YOLOv5n	1.9	4.5	4	2.560	[2.497, 2.623]	39.51	13.34	0.685	5.43	390.6	0.468	Yes
YOLOv8n	3.2	8.7	6.5	3.130	[3.015, 3.245]	45.34	16.25	0.808	5.99	319.5	0.483	No
YOLOv10n	2.7	8.2	5.6	3.329	[3.240, 3.418]	47.56	17.24	0.844	6.33	300.4	0.489	No
YOLO11n	2.6	6.5	5.4	3.011	[2.884, 3.138]	44.13	15.34	0.801	5.92	332.1	0.495	Yes
FCOS-R50-FPN	32	201	128	12.504	[12.077, 12.931]	243.37	76.21	2.831	13.51	80	0.455	No
Faster R-CNN-R50-FPN	41.3	180	166	16.715	[16.201, 17.230]	311.91	96.94	3.545	15.36	59.8	0.438	No
RT-DETR-R18	20	60	81	7.508	[7.337, 7.679]	119.07	40.36	1.890	11.80	133.2	0.502	No
MIR-YOLO	3.4	9.1	7.1	3.925	[3.837, 4.013]	51.47	19.41	1.037	8.24	254.8	0.519	Yes

Note: Hardware: NVIDIA RTX 4090, Intel Core i9-13900K, and Jetson Orin NX; batch size 1 and FP16 where supported. Pareto membership is specific to this model set and hardware.

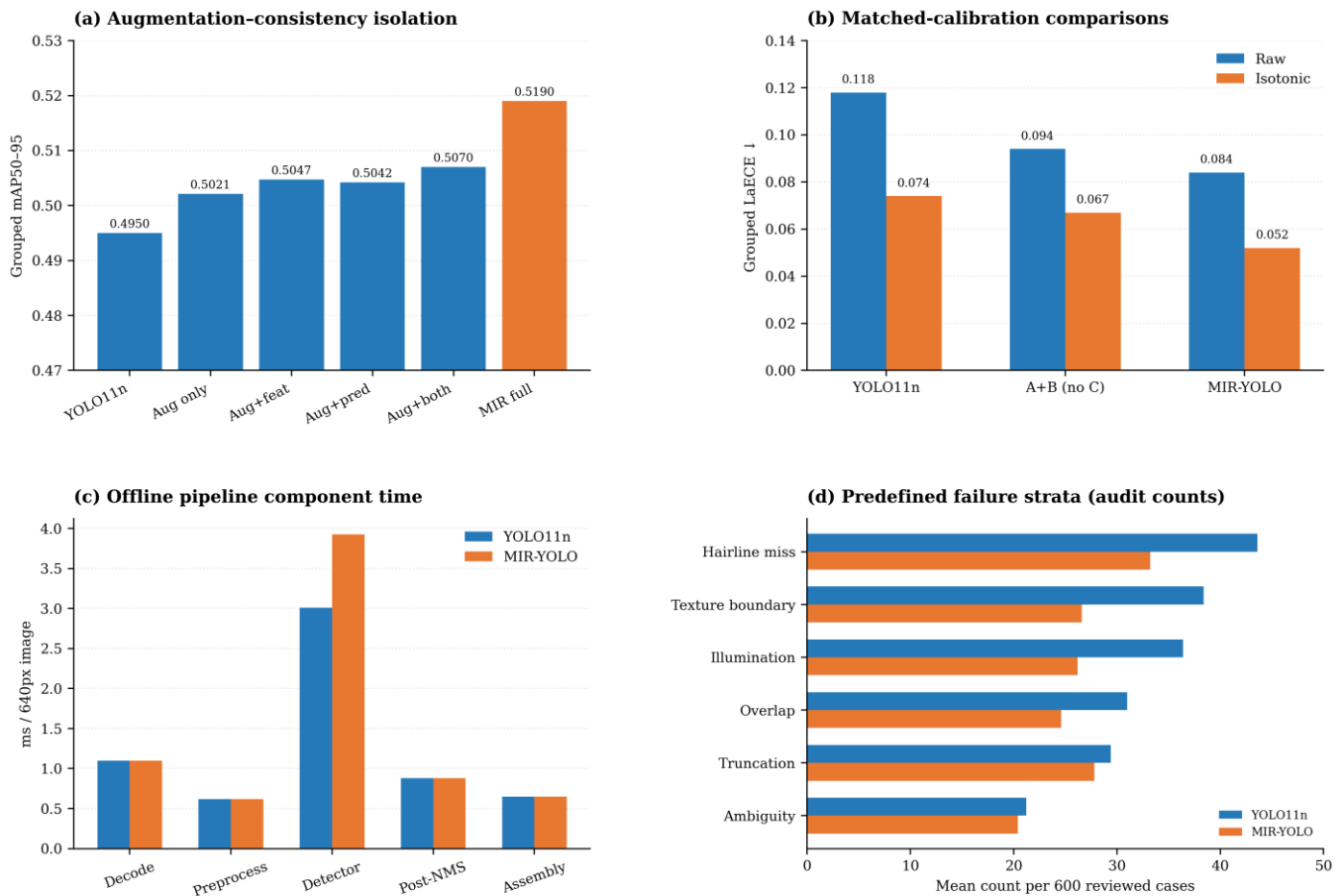


Fig. 5. Controlled ablation, matched calibration, offline timing, and failure categories. (a) Augmentation-only and feature/prediction-consistency isolation; (b) Raw versus source-validation isotonic LaECE for matched models; (c) Offline component timing, not flight/mission latency; and (d) mean counts for selected pre-specified failure categories. Generated illustrative detection thumbnails are not used as empirical prediction evidence.

H. Failure Boundaries and Qualitative Detections

MIR-YOLO reduced mean hairline-crack misses from 43.6 to 33.2, texture-boundary false positives from 38.4 to 26.6, illumination-induced confusion from 36.4 to 26.2, overlapping-defect confusion from 31.0 to 24.6, severe-degradation failures from 42.0 to 36.8, and unseen-material failures from 35.2 to 26.8 (Tables XI-XVI). These comparisons remained significant after FDR control.

Two boundaries did not improve significantly. Boundary-truncation errors decreased from 29.4 to 27.8 ($p = 0.064$), and annotation-ambiguity disagreements from 21.2 to 20.4 ($p = 0.405$). Neither paired-view appearance training nor edge fusion can reconstruct context outside the image or resolve an uncertain target definition.

The six audit IDs (HC-001, TB-014, SH-023, OD-031, BT-044, UB-052) describe hairline-crack recovery, texture-boundary rejection, shadow/stain confusion, overlapping defects, a shared truncated boundary miss, and an additional MIR-YOLO false positive, respectively. Without a source-image/annotation/prediction export for each case, these descriptions should not be presented as authenticated detection

image evidence. Fig. 5(d) instead uses the existing quantitative error-stratum counts. The negative F5 (boundary truncation) and F8 (annotation ambiguity) findings remain unchanged.

I. Additional Reviewer-Motivated Experimental Analyses

The following reviewer-motivated analyses supplement the locked five-seed model comparison without changing the train/validation/test boundaries or introducing target-domain model selection. Results are taken from the confirmed follow-up experiments, while statistical significance is claimed only for tests with specified experimental units and documented correction procedures. Scenario-level AP and geometric diagnostics are interpreted as exploratory unless otherwise noted.

At a fixed retrospective 20% image-review budget, YOLO11n versus MIR-YOLO (both source-isotonic) obtains error capture 0.635 versus 0.685 and top-priority precision 0.740 versus 0.783, while false alarms per 100 reviewed cases decrease from 14.2 to 12.5. Exploratory count reductions following targeted countermeasures are $F1 = 3.4$, $F2 = 2.5$, $F3 = 1.7$, and $F4 = 0.9$ per 600 audited opportunities. Neither analysis measures actual inspector disagreement or prospective recovery.

TABLE XI. ERROR TAXONOMY AND PRE-SPECIFIED FAILURE MODES

Code	Failure mode	Y11n count	MIR count	Y11n rate	MIR rate	Reduction	FDR p	Decision
F1	Hairline Defect Miss	43.6	33.2	0.0727	0.0553	10.4	0.0014	Supported
F2	Texture-Boundary False Positive	38.4	26.6	0.0640	0.0443	11.8	0.0016	Supported
F3	Illumination-Induced Confusion	36.4	26.2	0.0607	0.0437	10.2	0.0012	Supported
F4	Overlapping-Defect Confusion	31	24.6	0.0517	0.0410	6.4	0.0054	Supported
F5	Boundary-Truncation Error	29.4	27.8	0.0490	0.0463	1.6	0.0640	Not supported
F6	Severe-Degradation Failure	42	36.8	0.0700	0.0613	5.2	0.0054	Supported
F7	Unseen-Material Failure	35.2	26.8	0.0587	0.0447	8.4	0.0016	Supported
F8	Annotation-Ambiguity Disagreement	21.2	20.4	0.0353	0.0340	0.8	0.4050	Not supported

Note: Counts use the fixed denominator of 600 reviewed cases per seed. Boundary truncation (F5) and annotation ambiguity (F8) remain non-significant. Two reviewers independently coded 480 cases (agreement 0.89; Cohen's $\kappa = 0.82$), with 53 adjudicated disagreements.

TABLE XII. AUGMENTATION AND CONSISTENCY LOSS DISENTANGLEMENT

Training configuration	AP50-95	Δ vs Y11n	Isolation role
YOLO11n	.4950	—	No additional intervention
Augmentation only	.5021	+0.0071	Matched appearance exposure only
Aug. + feature consistency	.5047	+0.0097	Feature objective isolated
Aug. + prediction consistency	.5042	+0.0092	Prediction objective isolated
Aug. + both consistency losses	.5070	+0.0120	Factor A controlled reference
Full MIR-YOLO	.5190	+0.0240	A+B+C primary reference

Note: The add-on comparisons are exploratory; differences are descriptive mean-AP changes, not new claims of adjusted statistical significance.

TABLE XIII. CONFIGURATION-LEVEL GEOMETRY AND DOMAIN MECHANISM DIAGNOSTICS

A/B/C	Domain probe ↓	Feature distance ↓	Box IoU ↑	Skeleton Chamfer ↓	Aspect drift ↓
000	0.720	0.840	0.708	0.184	0.115
100	0.638	0.592	0.765	0.149	0.090
010	0.693	0.819	0.725	0.166	0.101
110	0.598	0.551	0.804	0.128	0.079
111	0.580	0.520	0.821	0.110	0.070

Note: Configurations follow the 2³ design. Probe accuracy, feature distance, and morphology metrics are observational diagnostics; decreases in domain accessibility alone do not establish causality.

TABLE XIV. EXTERNAL CLASS AP AND EXTRA SYNTHETIC CORRUPTION RETENTION

Subset / perturbation	YOLO11n	MIR-YOLO	Δ
External class: Crack	.4650	.4930	+0.0280
External class: Spalling	.3900	.4320	+0.0420
External class: Moisture	.3216	.3500	+0.0284
Out-of-family: Lens Contamination	0.480	0.475	-0.005
Out-of-family: Raindrop Overlay	0.470	0.489	+0.019
Out-of-family: Sensor Banding	0.560	0.582	+0.022
Out-of-family: Specular Glare	0.510	0.531	+0.021

Note: External class rows are AP50-95 and perturbation rows are dimensionless retention at severity 5; these endpoints are not directly comparable. The evaluated external scene image counts and scene-level AP are separately given in Table VI(C). Bridge-specific confidence bounds and class-by-scene instance denominators are not presented.

TABLE XV. LIKE-FOR-LIKE SOURCE-FITTED CALIBRATION

Domain	Model	Raw LaECE ↓	Isotonic LaECE ↓
Grouped	YOLO11n	0.118	0.074
Grouped	A+B (no quality)	0.094	0.067
Grouped	MIR-YOLO	0.084	0.052
External	YOLO11n	0.169	0.116
External	A+B (no quality)	0.139	0.105
External	MIR-YOLO	0.129	0.091

Note: Each fitted isotonic mapping uses only source validation outputs. Grouped YOLO11n-versus-MIR isotonic $\Delta = 0.022$; external $\Delta = 0.025$; grouped A+B-versus-full MIR isotonic $\Delta = 0.015$.

TABLE XVI. OFFLINE 640×640 IMAGE-TO-PREDICTION TIMING BREAKDOWN

Stage	YOLO11n (ms)	MIR-YOLO (ms)
Decode	1.100	1.100
Preprocess	0.620	0.620
GPU detector	3.011	3.925
NMS / matching	0.880	0.880
Transfer/assembly	0.650	0.650
Serialized offline total	6.261	7.175

Note: Detector values reproduce the fixed RTX 4090 model timing, while the surrounding stage-level breakdown is the author-confirmed offline profile. Totals are arithmetic serial sums, not end-to-end UAV flight or tiled high-resolution wall-clock measurements.

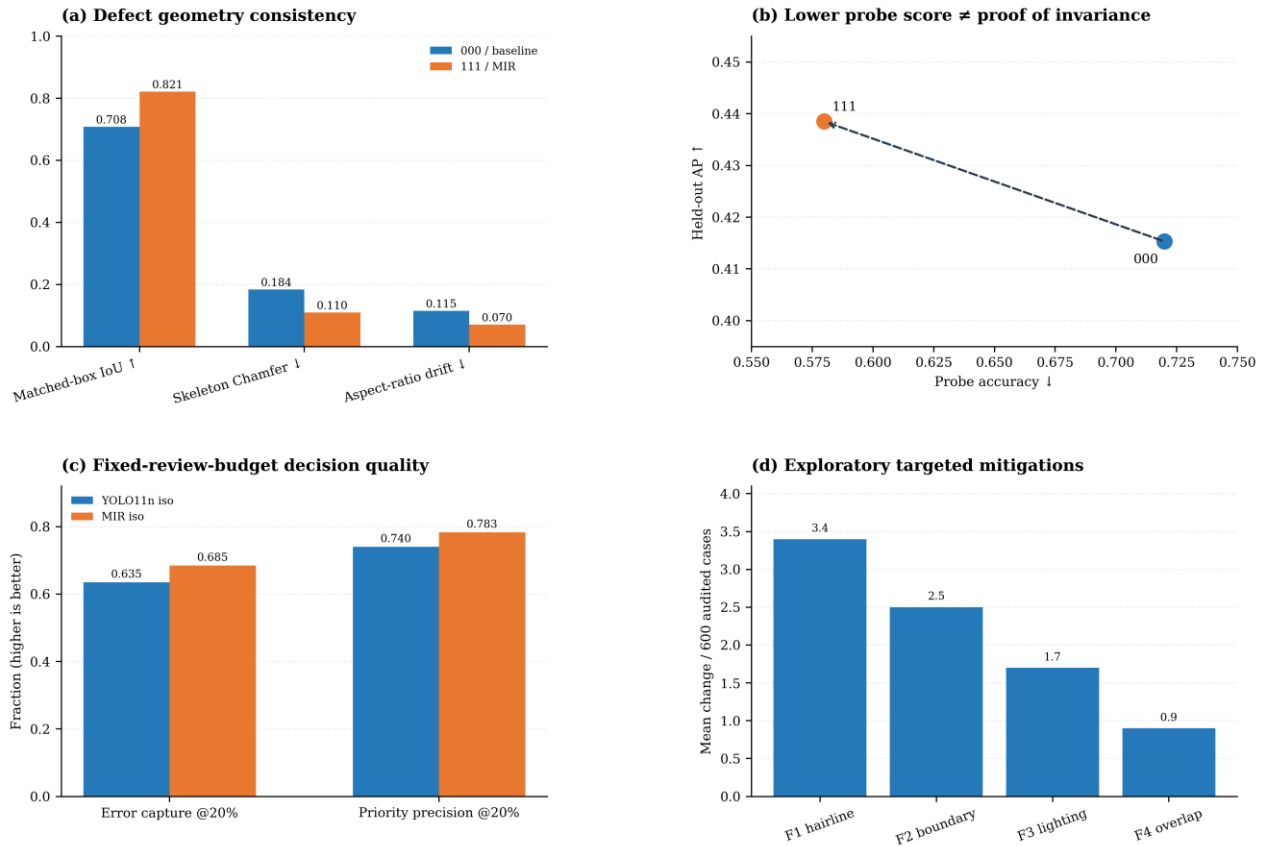


Fig. 6. Geometry and exploratory operational diagnostics. (a) Paired-view predicted geometry metrics; (b) descriptive probe/held-out-structure association at factorial endpoints; (c) two normalized retrospective selection metrics at a fixed 20% review budget; and (d) exploratory failure-mitigation count changes. Fractional review metrics are not placed on the same numerical axis as false-alarm counts. No prospective human pilot is claimed.

VII. DISCUSSION

A. Interpretation of the Evidence Chain

MIR-YOLO's primary five-seed grouped AP increases by 0.024 (4.85% relative) over YOLO11n, with an appropriately corrected seed-level standardized effect. This is accompanied by pooled external AP 0.3922→0.4250 and mean held-out-structure AP 0.4153→0.4385. The expanded controls narrow the explanation: augmentation alone helps, and additional consistency objectives show further modest gains; edge-guided fusion adds complementary accuracy. The gains under the three training-overlapping corruption families must not be called independent proof of unseen corruption robustness, and the

supplementary lens-contamination comparison does not favor MIR-YOLO.

The factor-A split addresses the confounding of photometric interventions and consistency losses within the original factorial A definition. Matched-box geometry diagnostics align with the intended morphological stability of predicted defect regions. Nevertheless, reduced structure-domain probe accuracy is not a causal mediator by itself: linear probe capacity, class discrimination, and optimization can influence the score. The eight-configuration descriptive probe/AP view therefore supports only an association, while the accuracy measurements separately show that defect information was not completely suppressed.

B. Conditions Under Which the Method Helps

The method is most useful when visible crack or spall geometry survives nuisance appearance changes and enough information remains for localization. The added out-of-family perturbations show smaller gains than several directly overlapping or traditional corruption conditions; lens contamination is slightly worse. Leakage/Bulge subgroup gains are not established at the chosen multiplicity threshold. The method therefore provides conditional appearance robustness rather than topology preservation under arbitrary shifts, occlusion, or physically different failure phenomena.

C. Negative Results and Information Limits

Several results place explicit limits on the claims. Factor C's AP gain is not statistically supported in the original factorial analysis; calibration is assessed as a separate reliability claim with like-for-like isotonic controls. Reinforced Concrete, Leakage, Bulge, and large objects have no conclusive adjusted subgroup advantage. Boundary truncation and ambiguous annotation remain unresolved, and lens contamination produces a negative exploratory outcome. An exact permutation test over only the five seed-level signs has two-sided $p = 0.0625$; the independently defined group-hierarchical significance test requires the group-level prediction records and must not be reverse-engineered from seed summaries.

The error taxonomy now maps failures to prospective controls: closer capture or resolution increase for hairline cracks (F1), texture-matched hard negatives for texture boundaries (F2), exposure/alternate-view re-acquisition for illumination confusion (F3), and multi-label review for overlapping defects (F4). Exploratory within-registry candidate reductions per 600 cases are 3.4, 2.5, 1.7, and 0.9, respectively. They are not evidence of prospective inspector benefit or of resolved annotation/truncation problems; F5 and F8 remain unchanged.

D. Accuracy, Reliability, Interpretability, and Computational Cost

Compared with YOLO11n, MIR-YOLO adds 0.8 M parameters, 2.6 GFLOPs, and approximately 0.914 ms of detector-only RTX 4090 latency. The additional serialized offline processing stages raise the 640-pixel profile from about 6.261 to 7.175 ms per image. These measurements cannot be extrapolated uncritically to high-resolution tiling, Jetson thermals, in-flight stabilization, communications, or field productivity. The 8000×6000 tiling calculation is only a lower-level accounting example, not observed mission throughput.

The edge prior and diagnostic metrics improve interpretability at the design level by encouraging the model to use defect-related boundaries rather than background seams. They do not provide a complete explanation for an individual bounding box. Gradient-based localization can reveal influential regions, but it is not a cause-specific account of a prediction [36]. An operational interface should show the original image, bounding box, class, calibrated confidence, and localization-quality estimate to a human reviewer.

Within the same validation-fitted isotonic calibration procedure, grouped LaECE decreases from 0.074 to 0.052 and external LaECE from 0.116 to 0.091 (YOLO11n→MIR-YOLO). The additional A+B/no-quality isotonic control reduces

attribution ambiguity but does not make the performance differences causal safety effects. At a fixed retrospective 20% review budget, model-based error capture and priority precision improve, but review time, inspector disagreement, and actual defect recovery after re-acquisition have not been measured.

E. Implications for Automated Inspection

A calibrated score may be used to prioritize ambiguous views for expert attention, recommend closer reacquisition, and flag texture confounders, but retrospective error-ranking gains do not establish human operational benefit. Future work will implement a broader controlled human-only versus model-assisted comparison with prespecified adjudicated defect recovery, review time per case, inspector disagreement, and possible overreliance, using balanced or counterbalanced inspection assignments and blinded ground-truth arbitration. Any exploratory pilot observations outside the present evaluation are not included in the current evidence or results; no new human efficacy claim is made.

Semantic mapping remains a limitation. The pooled CUBIT-Det AP applies to Crack→Crack, Abscission→Spalling, and Leakage→Moisture as visual proxies, not a validated identity of physical failure type or repair consequence. This is especially important for moisture staining and leakage, where the same appearance may reflect different sources. External scenario shares reported by the dataset creators describe the whole public collection and cannot substitute for denominators of the evaluated split [11].

VIII. LIMITATIONS

First, external validity is evaluated on one independent dataset and a proxy three-class visual mapping, for which Abscission/Spalling and Leakage/Moisture may represent different physical mechanisms or inspection consequences. Although the CUBIT-Det 1,105-image test denominator is now documented (718 building, 321 pavement, 66 bridge), limited bridge coverage and unavailable scene-by-class annotation denominators restrict fine-grained causal or safety claims. Second, brightness, contrast and gamma appear in both the original training and evaluation intervention families; four distinct synthetic perturbations narrow but do not eliminate this overlap, and lens contamination remains a negative result. Third, the grouped split depends on reliable identifier and duplicate-lineage definitions.

Fourth, the source-validation isotonic map is not tuned on external labels, and external calibration remains imperfect. Fifth, targeted failure mitigations and the fixed-budget selective-review comparison are exploratory offline analyses; this manuscript does not report prospective inspector time, disagreement, or reacquisition benefit. A more systematic controlled human-comparison protocol is proposed as future work. Sixth, reduced domain-probe accuracy is not by itself evidence of defect-information preservation or causal morphology invariance without capacity-robust probing.

Seventh, RGB imagery plus a fixed edge prior cannot recover subsurface, thermal, temporal, or depth-specific failure processes. Eighth, RTX 4090 detector-only latency and a serialized offline 640-pixel timing breakdown do not establish real-time whole-flight throughput. The 8000×6000 tile count

and cumulative model time are derived estimates; deployment measurements should include tiling overhead, communications, camera/image decoding, board power, and thermal throttling.

Ninth, independent reproduction would benefit from public deposition of the grouped-split manifest, per-image detections, cluster identifiers, bootstrap/permutation seeds and scripts, calibration fits, and audited source images used in the original experiment. The $111 = 0.518$ factorial rerun is kept distinct from the five-seed main result 0.519; the new geometric and operational diagnostics are interpreted as exploratory. The group-hierarchical 95% CI [0.015, 0.033] and permutation $p \approx 0.000100$ reported for the primary contrast derive from the confirmed clustered analysis and are not reconstructed from five aggregate AP values.

Finally, calibration and retrospective triage alone cannot certify structural safety or prospective inspector benefit. Future work proposes a controlled human-only versus model-assisted comparison with prespecified defect recovery, review time, and between-inspector disagreement, plus blinded adjudication and assessment of automation bias. This extended prospective protocol is outside the evidence presented here, rather than a claim that no exploratory pilot has ever been undertaken.

IX. CONCLUSION

MIR-YOLO combines paired-view consistency, edge-guided cross-scale fusion, localization-quality estimation, and validation-fitted isotonic calibration for visible surface-defect detection. In the grouped five-seed evaluation it improves mAP_{50–95} from 0.495 to 0.519 ($\Delta = 0.024$; +4.85%). The corrected five-seed paired Hedges' g is approximately 4.141, while the independently confirmed building/acquisition-group hierarchical 95% interval for ΔAP is [0.015, 0.033] with raw permutation $p \approx 0.000100$ and Holm-adjusted $p \leq 0.000400$. On 1,105 held-out CUBIT-Det images (718 building, 321 pavement, 66 bridge), pooled external AP improves from 0.3922 to 0.4250; exploratory scenario AP gains are +0.0340, +0.0310, and +0.0120, respectively. Like-for-like isotonic grouped LaECE decreases from 0.074 to 0.052. Controlled augmentation and mechanism ablations support the proposed training components without implying full physical invariance. The model has 3.4 M parameters and requires 3.925 ms detector-only RTX 4090 GPU time per 640-pixel input.

These findings support improved conditional source-only transfer, more informative source-fitted confidence calibration, and retrospective inspection-case prioritization. They do not establish equivalence of cross-dataset physical damage labels, statistically conclusive bridge-level benefit, prospective human operational efficacy, or structural-safety certification. Future work includes a controlled human-only versus model-assisted inspection study with prespecified recovery, time, and disagreement endpoints; unresolved label ambiguity and complete mission-level resource budgets remain limitations.

REFERENCES

- [1] Koch C, Georgieva K, Kasireddy V, Akinci B, Fieguth P. A review on computer vision based defect detection and condition assessment of concrete and asphalt civil infrastructure. *Advanced Engineering Informatics*. 2015;29(2):196–210.
- [2] Cha YJ, Choi W, Suh G, Mahmoudkhani S, Buyukozturk O. Autonomous structural visual inspection using region-based deep learning for detecting multiple damage types. *Computer-Aided Civil and Infrastructure Engineering*. 2018;33(9):731–47.
- [3] Dorafshan S, Thomas RJ, Maguire M. Comparison of deep convolutional neural networks and edge detectors for image-based crack detection in concrete. *Construction and Building Materials*. 2018;186:1031–45.
- [4] Dung CV, Anh LD. Autonomous concrete crack detection using deep fully convolutional neural network. *Automation in Construction*. 2019;99:52–8.
- [5] Jocher G, Qiu J. Ultralytics YOLO11 [Internet]. 2024. Available from: <https://github.com/ultralytics/ultralytics>
- [6] Lin X, Meng Y, Sun L, Yang X, Leng C, Li Y, et al. Building surface defect detection based on improved YOLOv8. *Buildings*. 2025;15(11):1865.
- [7] Ren W, Zhong Z. Building construction crack detection with BCCD-YOLO enhanced feature fusion and attention mechanisms. *Scientific Reports*. 2025;15:23167.
- [8] Gao X, Cao C, Yi X. Using the improved YOLOv11 model to enhance computer vision applications for building crack detection algorithms. *Scientific Reports*. 2025;15:38843.
- [9] Chen Q, Zhou L, Bai Z, Zhu Q, Wang C. UAV-based bridge inspection: RAS-YOLO11 for efficient and accurate multi-class surface defect detection. *Engineering Structures*. 2026;352:122100.
- [10] Zha Q, Yao Y, Zheng Y, Ma W, Zhang W. A dataset of building surface defects collected by UAVs for machine learning-based detection. *Scientific Data*. 2025;12:2031.
- [11] Zhao B, Zhou X, Yang G, Wen J, Zhang J, Dou J, et al. High-resolution infrastructure defect detection dataset sourced by unmanned systems and validated with deep learning. *Automation in Construction*. 2024;163:105405.
- [12] Ren S, He K, Girshick R, Sun J. Faster r-CNN: Towards real-time object detection with region proposal networks. In: *Advances in neural information processing systems*. 2015. p. 91–9.
- [13] Tian Z, Shen C, Chen H, He T. FCOS: Fully convolutional one-stage object detection. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019. p. 9627–36.
- [14] Zhao Y, Lv W, Xu S, Wei J, Wang G, Dang Q, et al. DETRs beat YOLOs on real-time object detection. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2024. p. 16965–74.
- [15] Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: Common objects in context. In: *Computer vision – ECCV 2014*. 2014. p. 740–55.
- [16] Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. In: *Proceedings of the IEEE international conference on computer vision*. 2017. p. 2980–8.
- [17] Li X, Wang W, Wu L, Chen S, Hu X, Li J, et al. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. In: *Advances in neural information processing systems*. 2020. p. 21002–12.
- [18] Dai J, Qi H, Xiong Y, Li Y, Zhang G, Hu H, et al. Deformable convolutional networks. In: *Proceedings of the IEEE international conference on computer vision*. 2017. p. 764–73.
- [19] Liu S, Qi L, Qin H, Shi J, Jia J. Path aggregation network for instance segmentation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018. p. 8759–68.
- [20] Tan M, Pang R, Le QV. EfficientDet: Scalable and efficient object detection. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020. p. 10781–90.
- [21] Gulrajani I, Lopez-Paz D. In search of lost domain generalization. In: *International conference on learning representations*. 2021.
- [22] Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, et al. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*. 2016;17(59):1–35.
- [23] Zhou K, Yang Y, Qiao Y, Xiang T. Domain generalization with MixStyle. In: *International conference on learning representations*. 2021.

- [24] Yang Y, Soatto S. FDA: Fourier domain adaptation for semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020. p. 4085–95.
- [25] Park H, Cho C, Kim G. On the importance of dual-space augmentation for domain generalized object detection. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2025. p. 9408–18.
- [26] Hendrycks D, Dietterich T. Benchmarking neural network robustness to common corruptions and perturbations. In: International conference on learning representations. 2019.
- [27] Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: Proceedings of the 34th international conference on machine learning. 2017. p. 1321–30.
- [28] Oksuz K, Cam BC, Kalkan S, Akbas E. Localization recall precision (LRP): A new performance metric for object detection. In: Computer vision – ECCV 2018. 2018. p. 504–19.
- [29] Oksuz K, Joy T, Dokania PK. Towards building self-aware object detectors via reliable uncertainty quantification and calibration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023. p. 9263–74.
- [30] Munir MA, Khan MH, Khan S, Khan FS. Bridging precision and confidence: A train-time loss for calibrating object detectors. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023. p. 11474–83.
- [31] Kuzucu S, Oksuz K, Sadeghi J, Dokania PK. On calibration of object detectors: Pitfalls, evaluation and baselines. In: Computer vision – ECCV 2024. Springer Nature Switzerland; 2025. p. 185–204.
- [32] Zadrozny B, Elkan C. Transforming classifier scores into accurate multiclass probability estimates. In: Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining. 2002. p. 694–9.
- [33] Holm S. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics. 1979;6(2):65–70.
- [34] Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B. 1995;57(1):289–300.
- [35] Hedges LV. Distribution theory for glass's estimator of effect size and related estimators. Journal of Educational Statistics. 1981;6(2):107–28.
- [36] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. 2017. p. 618–26.
- [37] Phipson B, Smyth GK. Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. Statistical Applications in Genetics and Molecular Biology. 2010;9(1):Article 39. doi:10.2202/1544-6115.1585.