

Road Surface Damage Detection Using Deep Learning: A Reproducible Framework for Comparing Detectors and Image Preprocessing

Amandeep Kaur^{1*}, Gurkamal Singh² and Parbhat Bansal³

¹UIE, Department of CSE, Chandigarh University, Mohali, India.

²Student, BE-CSE, Chandigarh University, Mohali, India.

³Student, BE-CSE, Chandigarh University, Mohali, India.

*Corresponding author: Amandeepkaur2177@gmail.com

²gurkamalsinghv@gmail.com ³parbhat.bansal001@gmail.com

Abstract—A road-damage detector must localize defects without confusing pavement texture, repairs, or shadows with damage. Its usefulness also depends on processing the complete camera pipeline within an acceptable delay. This paper specifies an evaluation framework for those two requirements using YOLOv8n, YOLOv10n, and YOLO11n on RDD2022. Four preprocessing conditions separate the effects of Gaussian smoothing and local contrast enhancement. The design distinguishes matched training and testing from enhancement applied only at inference, prevents related images from crossing dataset partitions where grouping information permits, and adds a held-country evaluation. Accuracy is assessed by class and country alongside end-to-end latency, with paired comparisons across repeated training seeds. A deployment design separates individual frame detections from reviewed road-defect events. The contribution is a testable methodology and an explicit decision rule for selecting a detector under latency and recall constraints. No training logs, predictions, or measured results accompany this study; consequently, it makes no claim that a particular detector or preprocessing method performs best.

Index Terms—Road damage detection, RDD2022, YOLO, preprocessing, domain shift, reproducibility.

I. INTRODUCTION

A camera survey of a road produces observations, not maintenance decisions. A dark region may be a pothole, a shadow, or an old repair; a narrow line may be a crack or a marking. Even when a bounding box is correct, repeated views of that defect should not automatically become separate repair requests. These distinctions motivate an evaluation that connects image detection to the eventual inspection workflow.

Maeda et al. established a practical precedent for learning road-damage categories from smartphone imagery [1]. RDD2022 broadens the available benchmark to 47,420 images from Japan, India, the Czech Republic, Norway, the United

States, and China, with more than 55,000 annotated instances [2]. Dataset diversity supports comparative study, but does not itself establish that a detector will transfer reliably between countries or capture conditions.

This paper retains the original study's focus on compact YOLO detectors and image preprocessing. It makes the experimental question more specific: does a preprocessing operation improve localization and classification enough to justify its added delay, and does that conclusion hold outside the training distribution? A visually clearer image is not sufficient evidence of a better detector. Enhancement can change the texture and intensity patterns on which a pretrained network depends.

Three research questions organize the protocol. *RQ1*: How do the selected detectors compare when data partitions, training budget, and evaluation are controlled? *RQ2*: What are the separate and combined effects of smoothing and contrast enhancement? *RQ3*: Are the accuracy and latency trade-offs stable across damage classes and held-country testing?

The contribution is methodological: an auditable split procedure, a factorial preprocessing comparison, a separate test for inference-time distribution mismatch, and a deployment selection rule. These are proposed design choices, not experimentally validated innovations. The manuscript is a research framework rather than a completed empirical comparison.

II. EVIDENCE BASE AND SCOPE

A. Road-damage data

The smartphone-based work in [1] connects image acquisition with detection rather than treating road inspection only as a laboratory image-classification problem. RDD2022 supplies a common four-class detection task [2]. The dataset repository states that training images have XML annotations, whereas the public test release contains images only [3]. Therefore, a local experiment cannot compute labeled test metrics on that public test folder without obtaining its ground truth or using an authorized challenge evaluation service.

The present protocol instead reserves part of the labeled re-

TABLE I
EVIDENCE AND ITS ROLE IN THE PROPOSED STUDY

Source	Supported role and boundary
Maeda et al. [1]	Smartphone-based road-damage detection precedent; no transferred performance claim.
RDD2022 [2, 3]	Dataset definition and label availability; internal split required for local scoring.
YOLO sources [4, 5, 6]	Model identity and inference design; relative road-detection ranking remains unmeasured.
OpenCV [7]	Enhancement implementation; accuracy impact remains a hypothesis.

lease for internal testing. Its results must be described as internal held-out evaluation, not as official challenge performance. The total size of RDD2022 also must not be reported as the number of labeled training images. A release-specific inventory is required before any split counts can be stated.

B. Detector selection

YOLOv8 provides a documented compact detection baseline [4]. YOLOv10 introduces consistent dual assignments to support inference without conventional non-maximum suppression (NMS) [5]. YOLO11 offers another documented set of detection models, including a nano variant [6]. The proposed comparison uses YOLOv8n, YOLOv10n, and YOLO11n to keep the study's scope manageable; it does not claim these are the newest available architectures.

The suffix “n” is not a guarantee of equal computational capacity. Model parameters, arithmetic cost at the chosen resolution, and memory use must be recorded separately. Published benchmark speed is not substituted for measurement on the experiment's hardware. Architecture-specific postprocessing is retained and identified, because forcing identical NMS onto an NMS-free detector would change the method being evaluated.

C. Preprocessing as a hypothesis

OpenCV documents CLAHE as local histogram equalization with contrast limiting [7]. Gaussian smoothing may suppress noise but can also weaken narrow crack boundaries. Local equalization may reveal low-contrast structure while emphasizing pavement texture. These competing possibilities motivate an ablation rather than a fixed assumption that enhancement is beneficial.

III. PROBLEM DEFINITION AND DATA GOVERNANCE

A. Detection target

For image I , the detector produces a set of predictions

$$D(I) = \{(b_j, c_j, s_j)\}_{j=1}^K, \quad (1)$$

where b_j is a bounding box, c_j is one of four damage categories, and s_j is a confidence score. The score is used for ranking and thresholding; it is not assumed to be a calibrated probability of damage.

The mapping in Table II is fixed in the conversion script and class configuration. Bounding-box detection identifies visible

TABLE II
RDD2022 LABELS AND EXPLICIT TRAINING IDENTIFIERS

ID	Label	Target category
0	D00	Longitudinal crack
1	D10	Transverse crack
2	D20	Alligator crack
3	D40	Pothole

regions; it does not measure crack width, pothole depth, or structural severity. Such quantities would need additional labels, calibration, or sensing.

B. Annotation audit

The data manifest records image identifier, country, dimensions, annotation path, valid object count, and a content hash. The audit checks that every image can be decoded, each label is supported, and each box has positive area and lies within the image. Missing annotation files are investigated rather than silently treated as negative examples. A genuinely annotated image containing no target objects remains useful as a negative example.

Coordinate conversion must follow the release's actual convention. If XML coordinates are one-based and inclusive, they are first converted to zero-based half-open corners; already compatible coordinates are left unchanged. For width W and height H , normalized YOLO coordinates are then

$$x'_c = (x_{\min} + x_{\max})/(2W), \quad w' = (x_{\max} - x_{\min})/W, \quad (2)$$

$$y'_c = (y_{\min} + y_{\max})/(2H), \quad h' = (y_{\max} - y_{\min})/H. \quad (3)$$

A sample of converted labels is overlaid on the original images. Suspect annotations are logged, and any correction policy is applied before partitioning. Discarding difficult annotations after seeing test errors would bias the evaluation.

C. Internal partitions and leakage control

The proposed target allocation is 70% training, 15% validation, and 15% internal testing within the labeled pool. These are allocation targets, not observed dataset counts. A seeded group assignment approximates those ratios while retaining country and class coverage. Exact images and object counts are published after the audit.

When route, sequence, or acquisition identifiers exist, all related frames are assigned together. Exact duplicates are linked by hash; candidate near-duplicates are reviewed using a documented image-similarity procedure. Connected duplicate groups remain in one partition. If route metadata is absent, visual deduplication reduces obvious overlap but cannot establish route independence. That limitation must accompany the reported results.

Augmentation occurs only after the split. Derived versions of an image inherit its partition. The test manifest is frozen before model selection, and its images are not used to choose preprocessing parameters, confidence thresholds, epochs, or augmentation settings.

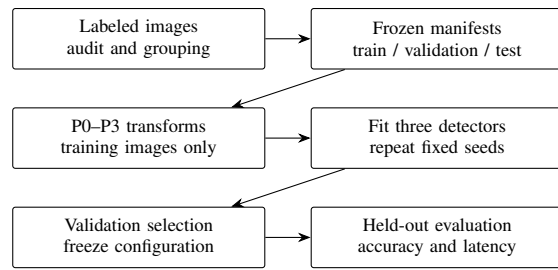


Fig. 1. Proposed experiment. Transformations are applied after partitioning; test data do not guide configuration selection.

D. Held-country evaluation

A secondary experiment reserves one country's labeled images exclusively for testing and trains on the remaining countries. Validation also uses only the remaining countries. India is a useful prespecified target for this project's local deployment interest; selecting a country because it gives favorable results is excluded. If resources permit, the procedure is repeated with each of the six countries held out in turn.

This test asks about transfer under a combined geographic and acquisition shift. It cannot isolate a purely national effect because camera setup, road texture, and collection practice may change together. Random pooled-split accuracy and held-country accuracy are therefore reported separately.

IV. PREPROCESSING AND MODEL PIPELINE

A. Operation order

The proposed pipeline decodes an image, optionally smooths and enhances it at its original resolution, converts the channel order expected by the detector, and then performs aspect-ratio-preserving resize with padding. Letterboxing to 640×640 avoids stretching crack geometry. The resize scale and padding are recorded so detections can be mapped back to original image coordinates.

B. Four-condition ablation

Table III defines a two-factor design. Smoothing is a 3×3 Gaussian filter with $\sigma = 0.5$ pixels and reflected boundary handling. CLAHE uses the 8-bit Lab luminance channel with a clip limit of 2.0 and an 8×8 tile grid; the other channels are retained before conversion back to color. These are prespecified starting settings, not tuned or validated optima. The OpenCV release and exact conversion constants are stored with the experiment.

The Gaussian kernel is

$$G(x, y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right). \quad (4)$$

P3 applies smoothing before luminance enhancement. An explicit CLAHE-only condition makes it possible to distinguish enhancement from the smoothing that preceded it in the original draft.

TABLE III
PROPOSED PREPROCESSING CONDITIONS

ID	Gaussian	CLAHE	Common final step
P0	Off	Off	Letterbox and normalization
P1	On	Off	Letterbox and normalization
P2	Off	On	Letterbox and normalization
P3	On	On	Letterbox and normalization

C. Matched training versus inference-only enhancement

In the primary experiment, a model trained under P2 is also validated and tested under P2; the same rule holds for every condition. This measures a complete training-and-inference choice. Optional geometric and photometric augmentation follows deterministic enhancement and is identical across conditions.

A separate mismatch experiment trains only the P0 baseline, then evaluates the frozen checkpoint under P1–P3. It answers whether preprocessing can be added to an existing detector without retraining. Its results are not pooled with the matched experiment. Otherwise, a deterioration caused by changed input statistics could be mistaken for evidence against preprocessing itself.

D. Training controls

The proposed primary budget is 100 epochs per run at 640×640 , using COCO-pretrained checkpoints and seeds 17, 29, and 43. The best validation mAP@0.50:0.95 checkpoint is selected within that budget. Early stopping is disabled for the main comparison so that each model receives the same maximum number of epochs. Checkpoint hashes identify the initialization precisely.

A training-only pilot determines the largest common physical batch size that fits the available accelerator, capped at 16. That value is frozen for all runs. SGD with initial learning rate 0.01, momentum 0.937, weight decay 0.0005, three warm-up epochs, and a cosine schedule ending at 1% of the initial rate is a proposed shared starting configuration. Any model that cannot use this recipe is reported as a separate training regime; silent reliance on different automatic optimizers is avoided.

Horizontal flipping at probability 0.5 is the primary geometric augmentation. Vertical flips and right-angle rotations are excluded because they change the road-view geometry and may change the meaning of directional labels. Mosaic, mixup, and random crops are disabled in the controlled baseline. Any richer augmentation is an additional experiment. These settings prioritize interpretability, and are not claimed to maximize each architecture's accuracy.

Three models, four preprocessing conditions, and three seeds produce 36 primary training runs. A reduced-budget study should reduce its declared scope before testing rather than omit unfavorable runs afterwards. Held-country training is an additional budget and is reported separately.

V. EVALUATION AND DECISION RULES

A. Detection metrics

A predicted box and a ground-truth box overlap according to

$$\text{IoU}(B_p, B_g) = \frac{|B_p \cap B_g|}{|B_p \cup B_g|}. \quad (5)$$

At a specified IoU and score threshold, predictions are matched to ground truth within the same class. Each ground-truth instance can match at most one prediction; duplicates count as false positives. Precision and recall are

$$P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{TP + FN}. \quad (6)$$

The evaluation reports mAP@0.50 and mAP@0.50:0.95, with the latter averaging AP over four classes and ten IoU thresholds from 0.50 to 0.95 in increments of 0.05. The evaluator version, interpolation rule, maximum detections, and handling of absent classes are fixed. AP needs a ranked set of predictions; it is not calculated only from a single deployment threshold.

Class-level precision, recall, and AP are accompanied by ground-truth counts. A class with no instances in a country slice is marked unavailable, not assigned a misleading zero. A pooled aggregate is accompanied by country-level scores so that a large source subset does not conceal a weak smaller subset.

B. Latency that reflects operation

Model-only latency and complete-pipeline latency answer different questions. For an unbatched serial implementation,

$$T_{\text{total}} = T_{\text{decode}} + T_{\text{prep}} + T_{\text{transfer}} + T_{\text{model}} + T_{\text{post}} + T_{\text{log}}. \quad (7)$$

The logging term includes creation of the output record; optional rendering and network upload are timed separately and labeled. Measurements use batch size one, 100 warm-up frames, and at least 1,000 measured frames or the entire test set if larger. GPU timing is synchronized where needed. All variants use the same image order, backend, precision, power configuration, and device.

Report the median and 95th-percentile latency, overall frames per second, and peak allocated accelerator memory. For N frames processed in elapsed time T , throughput is N/T ; it is not the average of individual reciprocal frame times. Timing includes online preprocessing even if enhanced images were cached for training. In asynchronous deployment, wall-clock latency and throughput are measured directly because overlapping stages do not follow the serial sum.

C. Paired ablations and uncertainty

For model m , preprocessing p , and seed s , define

$$\Delta_{m,p,s} = \text{mAP}_{m,p,s} - \text{mAP}_{m,0,s}. \quad (8)$$

Report the mean and standard deviation across the three paired seeds, keeping per-seed results available. For a fixed trained pair, a paired bootstrap resamples the same held-out groups and

recomputes the AP difference on each resample. A proposed 2,000 resamples yields a percentile interval for test-sample uncertainty. This interval is conditional on those checkpoints; it does not replace training-seed variability.

If only image-level groups are available, the interval is labeled accordingly and may understate dependence between nearby road views. Failed or unsupported bootstrap replicates are counted and disclosed. Claims about several exploratory comparisons are presented as exploratory rather than as multiple independent confirmations.

The interaction contrast

$$I_m = \text{mAP}_{m,3} - \text{mAP}_{m,1} - \text{mAP}_{m,2} + \text{mAP}_{m,0} \quad (9)$$

is calculated within each seed before aggregation. A positive observed contrast would suggest that the combined transformation exceeds the sum of the two isolated changes on this metric; it would not establish a universal benefit.

D. Selection under deployment constraints

Adding raw FPS to precision and mAP creates a score whose ranking depends on incompatible units. The proposed replacement is a constrained choice. Let ℓ_m^{95} be 95th-percentile end-to-end latency and $R_{m,D40}(\tau)$ pothole recall at threshold τ . On validation data, select

$$(m^*, p^*, \tau^*) = \arg \max_{m,p,\tau} \text{mAP}_{m,p} \quad (10)$$

subject to $\ell_{m,p}^{95} \leq L_{\text{max}}$ and $R_{m,p,D40}(\tau) \geq R_{\text{min}}$. The operating latency and recall requirements are supplied by the intended survey workflow before selection. Without agreed limits, report the nondominated accuracy-latency options instead of naming a deployment winner. If no configuration is feasible, report that outcome. The frozen choice is assessed once on the internal test set.

VI. ERROR ANALYSIS AND MONITORING DESIGN

A. Structured review of errors

Error review uses both a random sample and a separate sample of high-confidence mistakes, keeping their denominators distinct. Proposed categories include shadows, painted markings, repaired pavement, thin cracks, distant objects, blur, partial occlusion, localization error, and class confusion. Categories describe hypotheses to inspect, not findings already observed in this study.

Every reviewed case records the image identifier, prediction, reference box, confidence, preprocessing condition, and reviewer judgment. Two reviewers independently inspect a subset where practical and retain unresolved disagreements. Error counts are stratified by class and country. This helps distinguish a general enhancement effect from an improvement concentrated in a particular source subset.

Localization is examined separately from recognition. A box around a long, thin crack may include substantial intact pavement; its area is not the physical area of damage. Likewise, correctly recognizing a pothole with a poorly aligned box is

TABLE IV
REQUIRED EVIDENCE FOR THE EMPIRICAL VERSION

Artifact	Minimum content
Data manifest	Release, hashes, exclusions, group IDs, split membership, class and country counts.
Run manifest	Model and weight hash, code commit, package versions, seed, optimizer, augmentation, batch size.
Accuracy outputs	Per-image predictions, evaluator settings, per-class/country metrics, paired deltas.
Timing record	Device, OS, driver, backend, precision, warm-up policy, raw stage timings.
Audit examples	Original images and overlays, source IDs, error labels, reviewer decisions.

different from missing it entirely. Examples are shown at the original image scale with source identifiers once experiments exist.

B. From predictions to reviewed events

The proposed monitoring prototype accepts an image or a timestamped frame, applies the selected pipeline, and emits a detection record containing frame ID, original-coordinate box, class, score, model version, and preprocessing version. An operator can inspect the corresponding image before approving an event for maintenance review. This is an architectural proposal; a working prototype is not demonstrated by the supplied material.

Successive frames can contain the same defect. A temporal association module would link plausible repeated observations, but raw image IoU alone can fail as the vehicle moves. Its performance must be evaluated with event-level reference data before unique-defect counts are claimed. GPS can locate the camera at capture time; without geometric calibration it does not precisely locate each defect on the road surface.

C. Privacy and intended use

The research target is pavement damage. Images containing identifiable people or number plates should be redacted before external sharing, with access and retention controls for unredacted captures. Dataset terms are checked before redistribution. The system supports inspection triage; a damage box or confidence score alone does not establish an engineering repair priority.

VII. REPRODUCIBILITY AND REPORTING PLAN

The main results table in a completed study should contain one row per model and preprocessing pair, with mAP@0.50, mAP@0.50:0.95, pothole recall at the frozen operating threshold, latency percentiles, throughput, and memory. Seed variability belongs beside accuracy values. Class-wise and country-wise tables complement the aggregate comparison. Matched-training results, inference-only transformations, and held-country results remain in separate tables.

No numerical results table is supplied here because no runs are available. In particular, proposed hyperparameters are not represented as settings used in completed experiments. The artifact list in Table IV describes the evidence needed to move from the present protocol to a reproducible empirical paper.

VIII. LIMITATIONS AND NEXT STEPS

An internal split from a benchmark is also narrower evidence than a prospective survey. Duplicate grouping may miss related scenes, country labels confound multiple acquisition factors, and bounding-box annotations cannot support severity estimation. Three seeds provide a practical initial check on training variability, not exhaustive characterization. A shared optimization recipe may favor one architecture, so a later model-specific tuning study should use equal tuning budgets and report itself separately.

The next stage is to audit the labeled release, freeze manifests, execute the 36-run primary matrix, and publish predictions and timings. A held-India experiment then tests the local transfer question. Only after these stages should a selected detector be assessed on newly collected road video with manually reviewed event labels. Segmentation, calibrated depth measurement, and geospatial mapping are subsequent tasks with their own ground-truth requirements.

IX. CONCLUSION

This paper specifies how to compare compact road-damage detectors without conflating model choice, enhancement, data leakage, and runtime measurement. Its four-condition preprocessing design identifies individual and combined transformation effects, while a separate inference-only test examines input-distribution mismatch. Group-aware internal partitions and held-country evaluation address different forms of generalization. Deployment selection uses explicit latency and recall constraints rather than an arbitrary sum of metrics. The outcome is an executable research protocol whose conclusions remain open until measurements are obtained.

REFERENCES

- [1] H. Maeda, Y. Sekimoto, T. Seto, T. Kashiya, and H. Omata, "Road damage detection and classification using deep neural networks with smartphone images," *Computer-Aided Civil and Infrastructure Engineering*, vol. 33, no. 12, pp. 1127–1141, 2018, doi: 10.1111/mice.12387.
- [2] D. Arya, H. Maeda, S. K. Ghosh, D. Toshniwal, and Y. Sekimoto, "RDD2022: A multi-national image dataset for automatic road damage detection," *Geoscience Data Journal*, vol. 11, pp. 846–862, 2024, doi: 10.1002/gdj3.260.
- [3] Sekimoto Laboratory, "RoadDamageDetector: RDD2022 dataset and release documentation." Accessed: Sep. 9, 2026. [Online]. Available: <https://github.com/sekilab/RoadDamageDetector>
- [4] Ultralytics, "Ultralytics YOLOv8." Accessed: Sep. 9, 2026. [Online]. Available: <https://docs.ultralytics.com/models/yolov8/>
- [5] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding, "YOLOv10: Real-time end-to-end object detection," *arXiv preprint arXiv:2405.14458*, 2024, doi: 10.48550/arXiv.2405.14458.
- [6] Ultralytics, "Ultralytics YOLO11." Accessed: Sep. 9, 2026. [Online]. Available: <https://docs.ultralytics.com/models/yolov11/>
- [7] OpenCV, "Histogram equalization," *OpenCV 4.13.0 Documentation*. Accessed: Sep. 9, 2026. [Online]. Available: https://docs.opencv.org/4.13.0/d5/daf/tutorial_py_histogram_equalization.html