DEEP LEARNING

Deep Learning-Based Road Damage Detection Using UAV Imagery: A Review of Detection Methods, Geospatial Localization, and Inspection Systems

Ravi Khile, Prof. Archana Nilesh Pawar

[1]Department of Artificial Intelligence and Machine Learning Engineering, Navsahyadri Group of Institutions, Pune, Maharashtra, India; Savitribai Phule Pune University, Maharashtra, India
[2]Guide, Department of Artificial Intelligence and Machine Learning Engineering, Navsahyadri Group of Institutions, Pune, Maharashtra, India

Published November 2026

Abstract

Road damage inspection is essential for transportation safety and repair planning, yet conventional surveys are labour-intensive and hard to repeat across large networks. Deep learning can identify pavement defects automatically, and unmanned aerial vehicles (UAVs) offer flexible image collection, but recognizing damage in an image does not establish where it lies. This review examines how road damage recognition, UAV image acquisition, photogrammetric reconstruction, geospatial localization, and Geographic Information System (GIS) integration fit into one inspection workflow. Methods: An exploratory literature review covered road damage datasets, classification, object detection, segmentation, UAV survey planning, and spatial localization; it was not a PRISMA-based systematic review, and reported metrics are interpreted within each study's own evaluation setting. Findings: Public datasets such as RDD2022 are useful benchmarks although not exclusively aerial. Differences in datasets, damage categories, and protocols limit direct comparison, and recognition metrics cannot establish geographic accuracy. Key challenges are small defects, generalization, validated localization uncertainty, computational cost, and end-to-end field evaluation. Conclusions: Reliable systems should combine visual recognition, geographically validated mapping, traceable records, and human verification, with recognition and spatial accuracy evaluated separately.

Keywords: deep learning; geospatial localization; object detection; pavement inspection; road damage detection; UAV imagery

1 Introduction

Road infrastructure underpins economic activity, passenger travel, and the movement of goods, and keeping it serviceable depends on continuous pavement monitoring. Traffic loading, weather, poor drainage, construction weaknesses, and deferred maintenance gradually produce potholes, longitudinal and transverse cracks, fatigue cracking, and ravelling. Defects left undetected tend to worsen and drive up repair and rehabilitation costs, so systematic condition monitoring is central to infrastructure management.

Conventional inspection relies on field observation, manual measurement, photographic records, and occasionally specialized survey vehicles. These methods yield useful engineering data but demand skilled staff and time, are hard to repeat across large networks, and can vary with the inspector's experience, site conditions, and classification standards. Deep learning learns features directly from

annotated images, yet its results still depend on dataset quality and diversity, annotation consistency, model architecture, and how closely field conditions match the training data.

Automated analysis comprises several tasks. Classification labels a whole image or patch; object detection locates individual defects with bounding boxes; semantic segmentation labels every pixel; and instance segmentation also separates individual damaged regions. Faster R-CNN and YOLO are widely used detectors, while fully convolutional networks and U-Net derivatives are common for segmentation [1]–[4]. Progress also depends on annotated data: RDD2022 offers multinational road images for benchmarking [5], although geographic diversity within a dataset does not guarantee dependable results under different cameras, pavements, weather, or lighting.

UAVs add a flexible way to collect pavement imagery, covering long road sections from varied heights and angles and supporting repeated surveys where vehicle-based collection is difficult. More importantly, finding damage in an aerial photograph does not by itself give a usable map coordinate. Localization may require camera calibration, position and orientation data, ground control points, orthomosaics, photogrammetric reconstruction, or coordinate transformation; UAV photogrammetry has been used to build three-dimensional road models and estimate distress dimensions [6]. The reliability of the resulting spatial information depends on acquisition conditions, positioning accuracy, and validation.

A practical inspection framework therefore needs more than an accurate network. Image collection, preprocessing, recognition, localization, quality assessment, storage, visualization, and maintenance documentation all affect the final record: a pothole may be recognized yet mislocated, or precisely placed yet misclassified. Recognition accuracy and positional accuracy should be treated as separate but complementary measures. Earlier work has mapped trends in deep learning-based pavement distress detection [7]; this review instead links image-based recognition to the geographic usefulness of inspection results. It concentrates on visible surface defects and excludes subsurface diagnosis, structural health assessment, and autonomous repair.

The main contributions of this review are:

- A classification of deep learning approaches for road damage recognition by task: classification, detection, and semantic and instance segmentation.
- An examination of road damage datasets and of how annotation reliability, class imbalance, and domain differences affect training and evaluation.
- A discussion of UAV image acquisition, photogrammetric processing, and geographic localization, including the main sources of positional uncertainty.
- An integrated inspection framework linking UAV acquisition, deep learning, spatial processing, GIS visualization, and maintenance reporting.
- A summary of open challenges: generalization, small-defect recognition, validated geographic accuracy, efficiency, and field-level evaluation.

The following sections describe the review method, damage categories and datasets, recognition methods, UAV image acquisition, geotagging and localization, system integration and GIS reporting, a comparative analysis and research gaps, followed by the conclusion.

2 Review methodology

2.1 Design, scope, and search strategy

This is a targeted narrative review of deep learning-based road damage detection with UAV imagery, emphasizing detection methods, geospatial localization, datasets, and inspection systems. Studies using vehicle-mounted cameras, smartphones, and conventional road photographs are included when they supply datasets, benchmark methods, or techniques adaptable to UAV inspection, and results from different acquisition platforms are not assumed to be comparable. An exploratory search was run in October 2026 across publisher and scholarly records linked to IEEE Xplore, ScienceDirect, SpringerLink, Wiley Online Library, MDPI, Frontiers, and conference proceedings, using web-indexed records and publisher pages to locate papers and confirm metadata. A database-by-database execution log was not kept (see Synthesis framework and methodological limitations).

2.2 Selection criteria and evidence extraction

Studies were retained if they addressed damage detection or classification, UAV or aerial road inspection, documented datasets and benchmarks, detection, segmentation, or multi-scale architectures, localization and mapping of defects, or integration into condition-assessment systems. Numerical values are reported only when the source supports them and are read in the context of its own dataset, classes, resolution, split, and protocol, never as a controlled head-to-head comparison.

2.3 Synthesis framework and methodological limitations

The literature is synthesized along four dimensions: detection performance, UAV suitability, geospatial localization, and system integration from data collection to reporting. The search is exploratory. No Scopus or Web of Science export, database-level search log, duplicate-removal record, or formal screening log was produced, so the numbers of records retrieved, screened, and included cannot be reported as verified systematic-review figures. The review is therefore not a PRISMA systematic review and may be affected by selection and publication bias, differences in dataset composition, inconsistent evaluation protocols, and limited access to experimental detail.

3 Road damage categories and datasets

Datasets underpin training, validation, and evaluation, and road imagery varies with pavement, region, weather, lighting, camera, viewing angle, and severity, so dataset choice is a methodological decision.

3.1 Major categories of road damage

Longitudinal cracks run roughly parallel to traffic and range from faint hairlines to wide, irregular cracks, whereas transverse cracks run across the travel direction and can be confused in aerial images with lane boundaries, repair seams, and painted markings; annotation guidelines should therefore separate true defects from look-alikes.

Alligator (fatigue) cracking forms interconnected polygonal networks over a larger region and is hard to capture with one box; segmentation can outline it, although ambiguous boundaries cause annotator disagreement. Potholes are depressions whose appearance depends on depth, diameter, moisture, and lighting, and water, loose material, or shadow distort their boundaries. Detection suits

locating potholes for review and segmentation suits boundary or area estimation, but a single two-dimensional image does not give depth, which needs additional geometry.

3.2 Public road damage datasets

RDD2020 holds 26,336 road images from India, Japan, and the Czech Republic with more than 31,000 damage instances in four classes (longitudinal, transverse, and alligator cracks, and potholes), captured with vehicle-mounted smartphones [8]. A related study tested whether a model developed for Japan transfers to other countries and proposed models able to detect and classify damage in several countries [9], and the solutions submitted to the Global Road Damage Detection Challenge 2020 are summarized in [10]. RDD2022 extends the collection to 47,420 images from six countries (Japan, India, the Czech Republic, Norway, the United States, and China) with over 55,000 annotated instances in the same four classes [5]. According to the dataset preprint, its China data include a motorbike-collected subset and a drone-collected subset of 2,401 images with 3,068 labels [11]; RDD2022 therefore contains some aerial imagery but is not an exclusively aerial benchmark.

Ground-level cameras show individual defects in detail but not damage as seen from the air, where perspective, shadows, ground sampling distance, and background differ, so datasets built for conventional road imagery may need additional aerial samples or fine-tuning before they support UAV inspection reliably.

3.3 Dataset characteristics, annotation, and class imbalance

Beyond image count, dataset suitability depends on several properties:

- **Image diversity:** road materials, locations, lighting, weather, and pavement states.
- **Class distribution and label completeness:** enough examples per class, labelled consistently so that unlabelled defects are not treated as background.
- **Resolution and acquisition metadata:** whether narrow cracks survive resizing, plus camera parameters, altitude, location, time, and platform.
- **Training and test separation:** partitions that keep near-duplicate images out of both sets.

Annotation formats follow the task: image-level labels for classification, boxes and class identifiers for detection, pixel masks for semantic segmentation, and separated instances for instance segmentation. Class imbalance lets a model score well overall while failing on rare classes, so class-wise precision, recall, F1-score, and average precision give a fuller picture; remedies include targeted collection, class-aware sampling, augmentation, and re-weighted losses, none of which replaces genuinely diverse roads.

3.4 UAV suitability, data preparation, and evaluation

A defect's apparent size in a UAV image depends on altitude, focal length, sensor resolution, camera orientation, and surface geometry: higher flights cover more road but give small defects fewer pixels, while lower flights improve detail at the cost of more images. Ground sampling distance (GSD, Section V-B) relates pixels to ground distance, and datasets should report acquisition details beyond pixel dimensions. Two images taken a few metres apart in one flight are nearly identical, so a random split that separates them overestimates generalization; grouping by flight, road segment, date, or region is more demanding.

4 Deep learning methods for defect recognition

Deep learning learns visual features directly from images: early layers capture edges and textures, deeper layers represent cracks, pothole boundaries, and pavement edges, and these features feed classification, localization, or pixel-level delineation. Table 1 summarizes the four recognition tasks.

Table 1. Road damage recognition tasks.

Task	Output	Typical use and main limitation
Image classification	Label for an image or patch	Screening of large image sets. Gives no position within the image.
Object detection	Class, confidence score and bounding box per defect	Defect counting and approximate localization. Boxes around long, thin cracks include much background.
Semantic segmentation	Class label for every pixel	Extent and geometry of cracks or damaged area. Costly annotation; ambiguous boundaries.
Instance segmentation	Separate mask for each defect instance	Separate records for individual defects. Hard for overlapping or connected cracks.

4.1 Convolutional neural networks

Convolutional neural networks learn spatial patterns through filters, nonlinear activations, normalization, and downsampling, and usually serve as feature extractors for a classification layer, detection head, or segmentation decoder.

4.2 Object detection

Detection predicts a class, a confidence score, and a bounding box for each defect, giving more spatial information than classification at lower annotation cost than segmentation.

Two-stage detectors. Faster R-CNN proposes candidate regions and then classifies and refines them [1]. This supports accurate localization at the price of latency and suits offline processing, although boxes fit small cracks poorly and results depend on proposal quality and input resolution.

One-stage detectors. YOLO predicts classes and locations directly, without a proposal stage [2], and is widely considered for fast batch processing of aerial photographs or near-real-time screening.

Feature pyramids and small defects. Feature pyramid networks combine feature maps of different resolutions to handle objects of different scales [12], keeping spatial detail while retaining deeper semantics.

4.3 Semantic and instance segmentation

Segmentation predicts spatial masks and suits cases where damage shape and extent matter, at the cost of heavier annotation and post-processing and lower quality where boundaries are ambiguous or shadowed. Fully convolutional networks showed how convolutional architectures can produce dense predictions [3], and U-Net recovers spatial detail through an encoder–decoder with skip connections [4]; semantic segmentation can estimate affected area and crack-network geometry, provided image scale and distortion are handled. Instance segmentation assigns a separate mask to each object: Mask R-CNN adds a mask branch to a region-based detector [13], which helps when potholes must be counted or given separate geographic records, although overlapping or connected cracks make instance boundaries hard to define.

4.4 Transfer learning and fine-tuning

Transfer learning reuses features from a pretrained backbone, typically by replacing the output head, training it on road damage classes with the backbone frozen or updated at a small learning rate, and then fine-tuning. This is valuable when UAV-specific data are scarce, but pretrained features do not remove the need for representative target data, since models built on ground-level imagery may still need fine-tuning on aerial photographs.

4.5 Evaluation metrics

No single metric describes inspection reliability, so metrics should match the task. Precision is the share of predicted detections that are correct, and recall is the share of real defects that are found:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

where TP, FP, and FN are the numbers of true positives, false positives, and false negatives. False negatives leave defects unreported and false positives add to the manual verification workload, and their relative importance depends on the maintenance application. The F1-score is their harmonic mean:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

Intersection over union (IoU) measures the overlap between a predicted region and its reference region:

$$IoU = \frac{|B_p \cap B_g|}{|B_p \cup B_g|} \tag{4}$$

where B_p is the predicted region and B_g the reference region, computed on boxes for detection and on pixel masks for segmentation. Mean average precision (mAP) averages precision over classes at one or more IoU thresholds and should be compared only across compatible protocols; class-wise IoU and Dice scores complement detection metrics, and per-class results reveal weakness on rare classes that aggregate scores hide.

4.6 Model selection and limitations of current approaches

Model choice should follow the whole workflow rather than benchmark scores alone: the inspection objective (screening, in-image localization, or shape analysis), defect scale at the planned altitude, dataset compatibility, computational limits such as latency and onboard versus post-flight processing, localization requirements, the relative cost of missed defects and false alarms, and maintainability through reproducible evaluation.

Several limitations persist. Models are sensitive to distribution shift in road material, weather, camera, and processing. Small, low-contrast defects are easily missed, and higher resolution, multi-scale features, and tiling add cost and design choices. Benchmark metrics do not capture operational quality: a detector can find a defect yet give a wrong position, and a high segmentation score does not prove that estimated dimensions match physical measurements.

5 UAV-based inspection and image acquisition

The aerial viewpoint changes defect appearance, per-image coverage, and the geometric relationship between pixels and ground, so acquisition must be designed together with the detection model and localization method: a survey tuned for coverage may lack the detail for narrow cracks, while one tuned for detail needs more flight time and data.

5.1 UAV platforms, cameras, and flight planning

Multirotor UAVs hover and manoeuvre over selected segments, suiting localized inspection and repeated imaging, whereas fixed-wing UAVs cover larger areas efficiently but need different launch, landing, and planning arrangements. Altitude should be chosen from the smallest defect the system must recognize, not to maximize coverage, because higher flights cover more ground but can make small defects indistinguishable.

5.2 Ground sampling distance and image resolution

GSD is the approximate ground distance represented by one image pixel. For a simplified nadir-viewing camera over roughly planar ground it can be estimated as

$$GSD \approx \frac{H \cdot s}{f} \tag{5}$$

where H is the camera height above the ground, s is the physical pixel size on the sensor, and f is the focal length, in consistent units. Raising the camera increases GSD, so a narrow crack may span only a few pixels and become sensitive to compression, blur, contrast, and resizing.

5.3 Image quality control and imaging geometry

Reliable detection needs images that preserve defect appearance. Nadir images give broad surface coverage and simplify orthomosaic production, whereas oblique images reveal relief, pothole edges, and context but add perspective effects that complicate image-to-ground conversion; crack detection may favour consistent high-resolution nadir coverage, while three-dimensional analysis of potholes or deformation benefits from several viewpoints, at the price of more data.

5.4 Photogrammetry and orthomosaic generation

Photogrammetry uses overlapping images to estimate camera poses and scene geometry through feature extraction and matching, pose estimation, bundle adjustment, point-cloud generation, and, where required, surface reconstruction and orthomosaic production. Tan and Li reconstructed road three-dimensional models from UAV oblique images, detected pavement distress automatically, and extracted distress dimensions, reporting a height error of about 1 cm against a field survey for most cases [6].

5.5 Integration with deep learning, operations, and workflow

A typical workflow starts with acquisition and metadata collection; images are quality-checked and prepared, the model predicts in image coordinates, and predictions are tied to spatial information through the available geometric processing. A practical workflow has eight stages:

1. **Survey planning:** define the segment, objective, required detail, route, and operational constraints.

2. **Image acquisition:** capture imagery with suitable altitude, orientation, exposure, and overlap.
3. **Quality assurance:** identify blurred, incomplete, duplicated, or unsuitable images.
4. **Image preparation:** correct orientation, resize or tile, and keep the image–metadata link.
5. **Defect recognition:** apply a trained classification, detection, or segmentation model.
6. **Spatial processing:** transform predictions into geographic coordinates with a validated workflow when the data allow.
7. **Verification:** review uncertain predictions and assess recognition and localization errors.
8. **Reporting:** store verified records and present them as maps, tables, or maintenance reports.

6 Geotagging and geospatial localization

Maintenance staff need a defect's geographic position, not just its appearance in an image; until a prediction is tied to a ground position it remains an image-space observation. Reliable localization lets defects be mapped, assigned to road segments, compared across surveys, and prioritized.

6.1 Image coordinates, geographic coordinates, and geotagging

A detector returns pixel coordinates, such as the centre and size of a bounding box, or a mask of defect pixels; neither identifies a physical road location. A GIS stores positions in a spatial reference system, either latitude and longitude or projected easting and northing, and the chosen coordinate reference system (CRS) must be recorded so that records overlay correctly with road networks and orthomosaics. Many UAVs record coordinates, altitude, timestamp, and sometimes camera orientation at capture. This links an image to the surveyed area, but the camera position must not be treated as the position of every defect in the frame (Figure 1): a photograph covers considerable ground, and the offset grows with flight altitude and camera tilt. Geotagging therefore suits approximate survey location and organizing image sets, whereas defect-level localization needs additional geometry or a validated image-to-ground mapping.

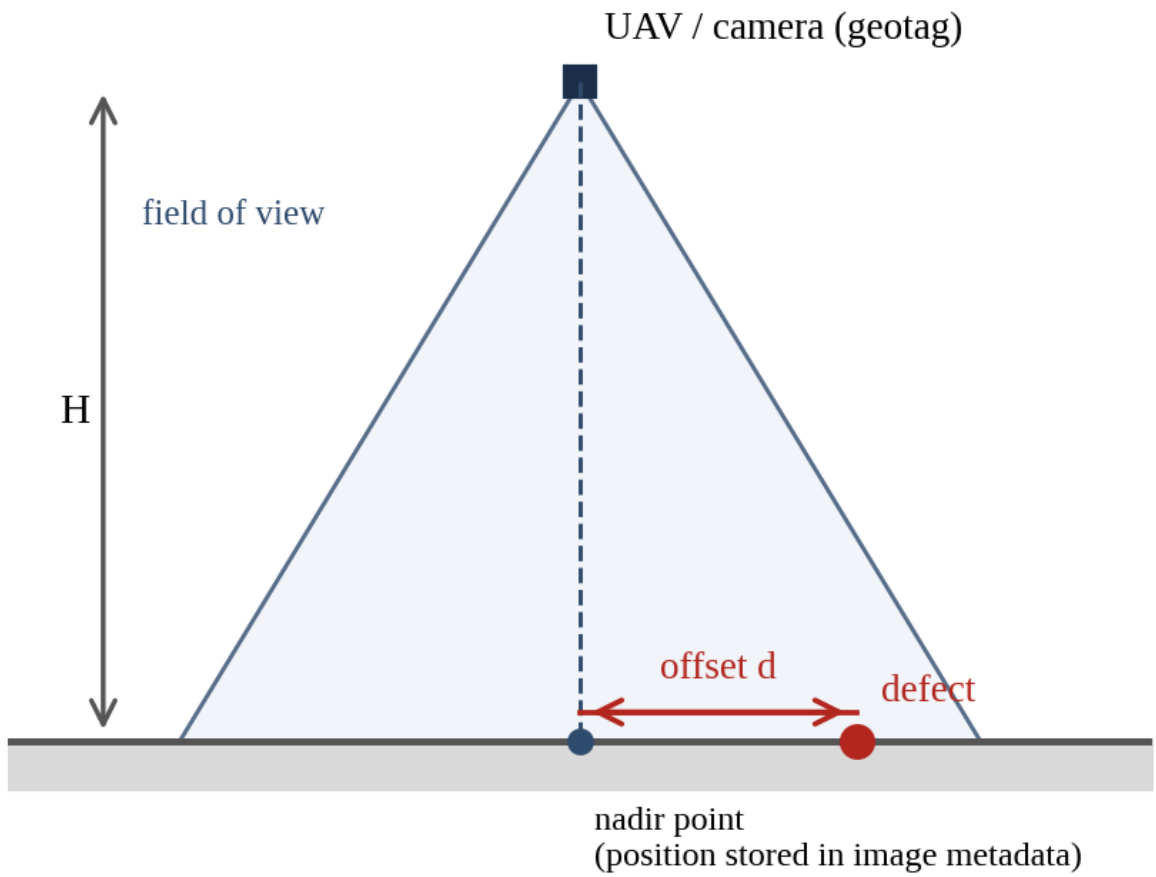


Figure 1. Camera position versus defect position. The geotag usually refers to the camera, so a defect lies at an offset d from the stored position that grows with flight height H and camera tilt (author's schematic).

6.2 Camera geometry, projection, and photogrammetric localization

In the pinhole model, a camera projects a three-dimensional point to image coordinates through its focal lengths and principal point (the intrinsic parameters), while the extrinsic parameters give its position and orientation relative to the world frame. For a locally planar road, a planar projective mapping (a homography) can relate image and ground coordinates when the camera geometry is known, but it becomes unreliable with elevation change, banking, or pothole depth, so the planar assumption should be stated. Ground control points (GCPs), surveyed locations with independent coordinates, can improve alignment to a known CRS, but their distribution matters, and independent checkpoints should be reserved for validation rather than used to fit the model.

6.3 Localization of detection and segmentation outputs

A detector returns a bounding box defined by its minimum and maximum image coordinates; the box centre is a convenient image-space reference point that can be projected to map coordinates once a valid transformation exists. A segmentation mask allows a representative point such as the centroid, or the whole boundary, to be transformed instead. Each localization record should preserve the image identifier, class, confidence, box or mask coordinates, transformation method, and resulting coordinates, so that a map marker can be traced back to its image evidence.

6.4 Coordinate reference systems and GIS compatibility

Geographic coordinate systems give angular positions, while projected systems give planar coordinates better suited to local distance and area. Image and road-network layers must be transformed consistently, with compatible definitions and units, before overlay or measurement, and each defect record should carry its CRS, source image, acquisition time, localization method, and any uncertainty estimate, because coordinates stored without a stated CRS become ambiguous across software.

6.5 Sources of localization error

Errors arise at several stages and can accumulate, so accurate recognition does not guarantee accurate positioning. The main sources are:

- **Positioning and orientation uncertainty:** errors in UAV position and camera attitude shift or misproject an entire image.
- **Calibration and altitude interpretation:** wrong intrinsic parameters, uncorrected distortion, or confusion between height above ground and above a reference surface add systematic offsets.
- **Surface geometry:** uneven pavement and three-dimensional defects violate planar projection assumptions.
- **Resolution and detection error:** low resolution blurs boundaries, and an incorrect box or mask displaces the defect's representative point.
- **Reconstruction error:** poor overlap, weak feature matching, and insufficient ground control degrade orthomosaic alignment.
- **Coordinate transformation:** wrong reference systems, units, or parameters cause additional displacement.

6.6 Validation of geographic accuracy and uncertainty

A localization method should be validated against reference positions independent of the data used to fit the transformation, obtained through surveying equipment, verified ground control, or another method with documented uncertainty. For N reference points, the horizontal error at point i is

$$e_i = \sqrt{\left(x_i^{ref} - x_i^{est}\right)^2 + \left(y_i^{ref} - y_i^{est}\right)^2} \tag{6}$$

where (x_i^{ref}, y_i^{ref}) are the reference coordinates and (x_i^{est}, y_i^{est}) the estimated coordinates in the same projected coordinate system. The root mean square error is

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} e_i^2} \tag{7}$$

RMSE summarizes the magnitude of horizontal error. It should be reported with the number and distribution of reference points, the coordinate system, the reference measurement method, and the evaluation procedure, and complemented by the median and an upper percentile when a few large errors skew the mean. A practical system should also not present every coordinate as equally reliable. High detector confidence supports the visual prediction only, not the geographic coordinate; quality flags can mark detections with poor image evidence, weak positioning, or an unreliable transformation for manual review, and validated error estimates can be reported with the location.

7 System integration and GIS-based reporting

An operational inspection system joins several components rather than relying on one model: imagery is collected and quality-checked, damage is recognized, predictions are tied to spatial information, and records are stored in a form that supports verification and maintenance planning.

7.1 Overall system architecture

A UAV-based system can be organized into six logical layers (Figure 2): data acquisition, image management, deep learning inference, geospatial processing, data storage, and presentation, running on one computer or distributed across devices and services. Stage outputs should stay separate, for example raw detector output apart from later computed geographic locations, so that the localization method can be revised without rerunning detection and errors can be attributed to recognition or spatial processing.

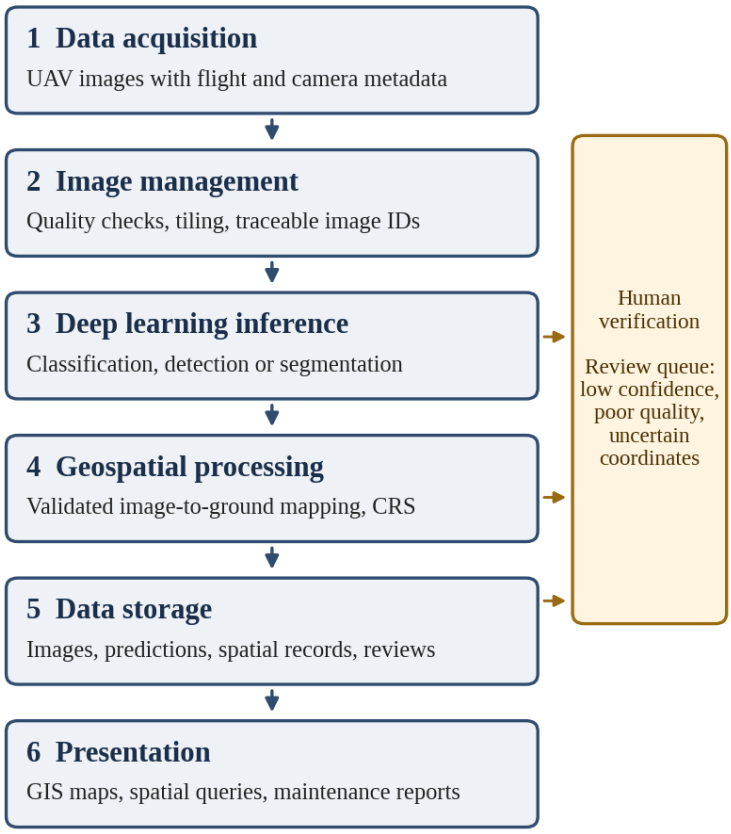


Figure 2. Six-layer architecture of a UAV-based road inspection system, with a human-verification loop for low-confidence, poor-quality or uncertain-coordinate records (author's schematic).

7.2 Image management and deep learning inference

Each image receives a unique identifier linked to its metadata (capture time, flight, positioning, and camera details), and originals are preserved because later resizing or compression can remove detail needed for verification. Quality-control decisions are recorded rather than discarded silently, and each prediction (class, confidence, box or mask, model identifier, source image) is kept in image coordinates.

7.3 Geospatial processing and database design

A conceptual schema can include mission records (identifier, date, road segment, platform, acquisition configuration), image records (file location, capture time, metadata, quality status), detection records (source image, class, confidence, image coordinates, model version), spatial records (geometry, CRS, localization method, uncertainty), verification records (reviewer decision, corrections, time, comments), and maintenance records (priority, status, follow-up). Separating these entities reduces duplication, lets several detections reference one image, supports repeated surveys, and preserves history; reviewer corrections should be stored apart from the original prediction, and the database should distinguish a predicted defect, a manually verified defect, and a completed maintenance action.

7.4 GIS visualization, reporting, and prioritization

A GIS shows defects in context with road centre lines, segment identifiers, administrative boundaries, and other infrastructure layers, as points for approximate locations, lines for elongated features, or polygons for validated footprints, with the representation matched to spatial data quality. Visualization should communicate uncertainty, for instance by drawing approximate coordinates differently from independently validated ones and letting users open the source image behind any feature; a map marker alone does not guarantee location accuracy.

7.5 Deployment, reliability, and data governance

Inference can run on the UAV, a nearby ground station, or a remote service. Onboard processing reduces transmission of raw images and may flag suspected defects during a mission, but payload, power, heat, memory, and endurance constrain the hardware, so large segmentation models may be unsuitable without optimization. A complete system should also handle failures and preserve record integrity: processing jobs may fail from corrupted files, memory shortage, or software errors and should be logged and resumed without duplicating results; each detection should identify the model and configuration that produced it; imagery and geographic records should be protected through authentication, permissions, secure transfer, backups, and retention policies; and governance should define who may review predictions, correct locations, approve records, and record maintenance, keeping automated predictions visibly distinct from verified observations.

7.6 End-to-end system workflow

Its effectiveness depends on each component and on the consistency of their interfaces: reliable recognition, validated localization, traceable records, and clear uncertainty reporting are all needed for repeatable and accountable infrastructure management.

8 Comparative analysis

8.1 Overview of the literature

Research has moved from conventional image processing, which relies on hand-designed features and thresholds and degrades under changing illumination, shadows, textures, and clutter, to deep learning-based detection and segmentation. RDD2020 and RDD2022 collect damage images across geographic and environmental conditions [5], [8], but they are not exclusively UAV-captured, so results on them are not automatic evidence of equal performance on aerial imagery.

8.2 Summary of selected studies

Table 2 summarizes representative studies on datasets, damage detection, pavement assessment, and UAV inspection, emphasizing task, data source, reported evidence, and limitations rather than ranking models. The numbers in Table 2 come from different datasets, class definitions, and protocols and must not be compared across rows.

Table 2. Comparison of selected road damage detection studies.

Study	Data or source	Method and task	Reported evidence	Limitations
Ren et al. (2015) [1]	General object-detection benchmarks	Faster R-CNN; region-based detection	Introduced region proposal networks for detection	Not road-damage-specific; proposal generation and multi-stage processing can increase inference cost.
Redmon et al. (2016) [2]	General object-detection benchmarks	YOLO; single-stage detection	Introduced a unified single-stage detector	Neither road-damage-specific nor UAV-specific.
He et al. (2017) [13]	General instance-segmentation benchmarks	Mask R-CNN; detection with instance segmentation	Added a mask-prediction branch to a region-based detector	Pixel-level masks need extra annotation effort and computation.
Tan and Li (2019) [6]	UAV oblique photogrammetric images	3D road model reconstruction; automatic distress detection and dimension extraction	Height error of about 1 cm against a field survey for most cases	Addresses mainly distortion-type distress; depends on image overlap, calibration and scene conditions.
Arya et al. (2020) [10]	GRDDC'2020 (RDD2020 data)	Overview of state-of-the-art challenge solutions	Benchmark-oriented comparison of detection solutions	Smartphone imagery; results depend on the benchmark data and protocol.
Arya et al. (2021) [8]	26,336 images from India, Japan and the Czech Republic; vehicle-mounted smartphones	Annotated dataset with four damage classes	More than 31,000 damage instances	Not UAV imagery; geographic and imaging differences affect generalization.
Arya et al. (2021) [9]	26,620 smartphone images from India, Japan and the Czech Republic	Deep learning detection and classification; cross-country assessment	Assessed use of a Japanese model in other countries; proposed multi-country models	Cross-country performance does not establish UAV performance.
Silva et al. (2023) [14]	UAV images; RDD2022 China subset combined with a Spanish road dataset	YOLOv4, YOLOv5 and YOLOv7 detection in UAV images	Reported as showing the potential of UAVs with deep learning for road damage detection (metrics not extracted here)	Combined data sources; results tied to the authors' protocol.
Arya et al. (2024) [5]	RDD2022: 47,420 images from six countries	Dataset and CRDDC'2022 benchmark; four damage classes	More than 55,000 instances; top challenge model F1 of 76.9% over the six countries	Mixed acquisition sources; not exclusively aerial.

Study	Data or source	Method and task	Reported evidence	Limitations
Cano-Ortiz et al. (2024) [16]	Road images; validation on RDD2022 and CPRI	YOLOv5 sub-models; detection, quantification and a geotagged assessment tool	mAP of 0.563 (RDD2022) and 0.570 (CPRI) for the refined architecture	Dataset-specific results; geotag-based rather than photogrammetric localization; validation not UAV-specific.
Zheng et al. (2024) [7]	Web of Science publications, 2016–2024	Review with CiteSpace bibliometric analysis	Maps hotspots and research fronts in deep learning-based distress detection	No new controlled detector comparison; search may not cover all research.
Yang et al. (2026) [15]	1,500 images from UAV-RDD and CrackForest (1,000/250/250 split)	YOLO-SCX, a YOLOv5n-based UAV road damage detector	mAP@0.5 of 66.3% and precision of 79.2%; gains of 5.8 and 6.0 points over the baseline	Small composite dataset and custom protocol limit comparison with other studies.

Values are reproduced as reported by the original authors under their own datasets and protocols and are not directly comparable across rows.

8.3 Comparison of architectures, datasets, and localization

Two-stage detectors such as Faster R-CNN can help when defects are small or localization quality matters, but they add inference time and suit offline processing. Single-stage detectors such as YOLO balance speed and accuracy on large image collections, although small, narrow, low-contrast cracks remain hard, and gains from multi-scale features, resolution choices, or careful annotation must be demonstrated on suitable test sets.

Detection and localization are related but distinct: a box or mask does not itself give a reliable coordinate. Among the studies in Table 2, one links geotagged images and detections in a software tool for road administrators [16], and one uses UAV photogrammetry to obtain three-dimensional distress dimensions [6].

8.4 Critical findings

The reviewed literature supports five conclusions.

1. Public road damage datasets support benchmarking, but acquisition conditions must be considered before results are transferred to UAV imagery.
2. Single-stage detectors are practical for efficient inference, whereas segmentation methods give more detailed damage geometry.
3. Detection scores alone do not establish geographic accuracy or the usefulness of a complete inspection system.
4. Cross-country and cross-platform evaluation helps assess generalization, provided evaluation protocols are documented clearly.
5. A deployable UAV inspection system requires integrated acquisition, inference, georeferencing, storage, visualization, and validation.

9 Research gaps and future directions

Deep learning has advanced road damage recognition, but UAV inspection is not yet reliable across diverse road networks. The challenges extend beyond network accuracy to dataset diversity, small defects, generalization, localization, efficiency, and integration with maintenance.

9.1 Dataset diversity, class imbalance, and annotation uncertainty

Public datasets may not cover every combination of pavement material, climate, road design, and platform, and aerial imagery adds variation in viewing angle, GSD, scale, and background, so a detector trained mostly on ground-level images may not transfer to high-altitude photographs.

9.2 Small defects and robustness in real flight conditions

Models built on carefully selected datasets may also degrade under wind, motion blur, changing light, shadow, cloud, wet pavement, and constrained flight altitude, so robustness evaluation should use realistic imagery with documented platform, camera, altitude, overlap, weather, and processing, and systems should flag unsuitable images and uncertain predictions using confidence thresholds validated for the target environment.

9.3 Joint evaluation of recognition and localization, and defect measurement

Future work should assess recognition and localization in one framework in which each mapped defect keeps its source image, image-space prediction, transformation method, CRS, and validation information, and independent ground-truth coordinates measure spatial error; the effect of detection errors on the representative point should be separated from geometric transformation error. Detection also does not establish physical dimensions or severity: neither a box nor a mask measures area or depth without a geometric relationship between image and road, and a two-dimensional RGB image may show a dark region without revealing pothole depth.

9.4 Efficiency, standardization, and end-to-end field validation

High-resolution UAV imagery needs considerable storage, memory, and time, so studies should report inference latency, model size, memory, and hardware, and separate inference time from total processing time, which also includes transfer, tiling, spatial transformation, and database updates. Differences in datasets, training, resolution, annotations, and metrics make published scores hard to compare, so standardized protocols should define defect categories, partitioning, matching criteria, and thresholds and extend to spatial accuracy and robustness. Finally, complete workflows should be evaluated on real survey missions, documenting coverage, processing time, recognition performance, geographic accuracy, verification effort, and operational failures.

9.5 Explainability, human verification, and temporal monitoring

Human review stays essential for uncertain predictions and poor imagery, so review interfaces should let users inspect the source image, correct the class or location, and record the reason, with versioned corrections feeding later model development. Repeated surveys support monitoring, but differences in angle, light, resolution, or localization can make a defect appear to change. Research should address defect association across surveys, separate visual change from physical change, and link observations to maintenance records, without marking a defect as repaired merely because a later image does not show it, since that may reflect a missed detection or inadequate coverage.

9.6 Future research directions and proposed research framework

Based on these gaps, the following directions are important for more reliable UAV-based road damage inspection:

1. Geographically diverse, well-documented aerial datasets and domain generalization to unseen environments and UAV configurations.
2. Small-defect recognition through multi-scale features, high-resolution processing, and adaptive tiling, with measured computational cost.
3. Joint recognition and localization evaluation, and reliable geometric measurement using calibrated imaging or three-dimensional information.
4. Efficient deployment using lightweight models, compression, and hybrid processing on representative hardware.
5. Standardized benchmarking with consistent metrics, partitions, hardware information, and spatial validation procedures.
6. Human-in-the-loop review, longitudinal evaluation of repeated surveys, and end-to-end field trials under operational conditions.

An integrated framework would treat inspection as connected but separately evaluated tasks: a documented UAV acquisition plan that retains the metadata needed to judge image quality and spatial reliability; a deep learning model that proposes candidate defects; and a geometric stage that assigns locations.

10 Conclusion

This review examined deep learning-based road damage detection with UAV imagery, covering recognition methods, aerial image acquisition, geospatial localization, and integrated inspection systems. Classification, object detection, and semantic or instance segmentation, illustrated by Faster R-CNN, YOLO, FCN, U-Net, and Mask R-CNN [1]–[4], [13], serve different needs, and the suitable choice depends on the objective, defect size, annotation availability, computing budget, and spatial detail required.

UAV effectiveness depends on flight planning, camera configuration, GSD, overlap, and operating conditions, and photogrammetry and orthomosaics provide mapping geometry whose quality depends on image matching, calibration, positioning, and georeferencing [6]. Localization is a separate requirement from recognition and should be validated against independent reference coordinates, with each record preserving its CRS, source image, transformation method, and uncertainty; predicted and verified defects must stay distinct, and model confidence must not be read as geographic accuracy or engineering severity. In conclusion, deep learning-based road damage detection combined with UAV imagery and geospatial processing is a promising route to automated infrastructure inspection, whose practical value rests on locating defects reliably, communicating uncertainty, preserving traceable records, and supporting informed maintenance within a reproducible workflow with human oversight.

Declarations

No external funding was received for this research. The author declares no competing interests. Ravi Khile was responsible for the conceptualization, literature review, analysis, and manuscript preparation. The author thanks Prof. Archana Nilesh Pawar for her guidance and academic support.

References

[1] 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, vol 28. 2015.

[2] Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: *Proc IEEE Conf Comput Vis Pattern Recognit (CVPR)*. 2016:779–788.

[3] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: *Proc IEEE Conf Comput Vis Pattern Recognit (CVPR)*. 2015:3431–3440.

[4] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. 2015:234–241.

[5] Arya D, Maeda H, Ghosh SK, Toshniwal D, Sekimoto Y. RDD2022: a multi-national image dataset for automatic road damage detection. *Geosci Data J*. 2024;11(4):846–862. https://doi.org/10.1002/gdj3.260

[6] Tan Y, Li Y. UAV photogrammetry-based 3D road distress detection. *ISPRS Int J Geo-Inf*. 2019;8(9):409. https://doi.org/10.3390/ijgi8090409

[7] Zheng L, Xiao J, Wang Y, Wu W, Chen Z, Yuan D, Jiang W. Deep learning-based intelligent detection of pavement distress. *Autom Constr*. 2024;168:105772. https://doi.org/10.1016/j.autcon.2024.105772

[8] Arya D, Maeda H, Ghosh SK, Toshniwal D, Sekimoto Y. RDD2020: an annotated image dataset for automatic road damage detection using deep learning. *Data in Brief*. 2021;36:107133. https://doi.org/10.1016/j.dib.2021.107133

[9] Arya D, Maeda H, Ghosh SK, Toshniwal D, Mraz A, Kashiyama T, Sekimoto Y. Deep learning-based road damage detection and classification for multiple countries. *Autom Constr*. 2021;132:103935. https://doi.org/10.1016/j.autcon.2021.103935

[10] Arya D, Maeda H, Ghosh SK, Toshniwal D, Omata H, Kashiyama T, Sekimoto Y. Global road damage detection: state-of-the-art solutions. In: *Proc IEEE Int Conf Big Data*. 2020:5533–5539. https://doi.org/10.1109/BigData50022.2020.9377790

[11] Arya D, Maeda H, Ghosh SK, Toshniwal D, Sekimoto Y. RDD2022: a multi-national image dataset for automatic road damage detection. *arXiv* [Preprint]. 2022 Sep. arXiv:2209.08538. https://doi.org/10.48550/arXiv.2209.08538

[12] Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: *Proc IEEE Conf Comput Vis Pattern Recognit (CVPR)*. 2017:2117–2125. https://doi.org/10.1109/CVPR.2017.106

[13] He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In: Proc IEEE Int Conf Comput Vis (ICCV). 2017:2961–2969.

[14] Silva LA, et al. Automated road damage detection using UAV images and deep learning techniques. *IEEE Access*. 2023;11:62918–62931. https://doi.org/10.1109/ACCESS.2023.3287770

[15] Yang H, Ma Z, Zhu H, Yu Q. *Road damage detection method based on UAV imagery and YOLO-SCX. Front Built Environ*. 2026. https://doi.org/10.3389/fbuil.2026.1681691

[16] Cano-Ortiz S, Lloret Iglesias L, Ruiz del Árbol PM, Lastra-González P, Castro-Fresno D. An end-to-end computer vision system based on deep learning for pavement distress detection and quantification. *Constr Build Mater*. 2024;416:135036. https://doi.org/10.1016/j.conbuildmat.2024.135036