Int. J. Sci. R. Tech., 2026 3(7)
A Multidisciplinary peer-reviewed Journal
www.ijsrtjournal.com [ISSN: 2394-7063]

Original Article

Simplifying Pavement Crack Detection: From Deep Learning Pipelines To Multimodal Vision-Language Models

Alfred Patrick Patric*

Independent Researcher

ABSTRACT

Pavement distress detection and pavement condition index (PCI) assessment are central to transportation asset management and road safety. Automated monitoring has progressed from handcrafted image-processing methods to convolutional neural-network detectors and semantic-segmentation networks. Although these models can achieve high detection accuracy, operational systems commonly require separate components for localization, segmentation, classification, severity estimation, and report generation. This narrative review synthesizes literature on deep-learning crack detection, generative augmentation, sensing platforms, localization, and pavement assessment, and compares conventional pipelines with multimodal vision large language models (VLLMs). The review introduces the Emmaus Pipeline, a decentralized framework in which high-definition cameras and GPS loggers mounted on routine public-service vehicles capture road video for depot upload and offline VLLM analysis. A unified model can produce crack localization, morphology, severity reasoning, and structured outputs suitable for asset-management databases, while GPS interpolation and visual odometry can improve geolocation. The comparison indicates that VLLMs may reduce pipeline complexity and labeling dependence, but they remain limited by computational demand, measurement precision, reproducibility, and the need for engineering validation. Future work should benchmark VLLMs against dedicated detectors using PCI-relevant metrics, multisensor synchronization, and longitudinal field studies. The proposed architecture offers a practical research direction for lower-cost, network-scale pavement monitoring.

Keywords: Pavement crack detection; deep learning; vision large language models; pavement condition index; multimodal sensing; road asset management.

INTRODUCTION

The structural integrity of road pavement directly influences traffic safety, vehicle operating costs, and economic efficiency. Traditional condition monitoring relies heavily on manual inspection, which is slow, subjective, hazardous for field personnel, and difficult to sustain across an entire local road network. Automated pavement-distress detection therefore has become an important component of modern road asset management [1,2].

Early automation combined computer vision with handcrafted image-processing techniques such as edge detection, thresholding, and morphological filtering. These methods are computationally lightweight but often fail under real-world variation in shadows, illumination, oil stains, lane markings, debris, and heterogeneous pavement textures.

The next wave of automation adopted supervised convolutional neural networks (CNNs). Object detectors locate visible distress, while semantic-segmentation networks estimate crack shape and area. In practice, however, a complete operational workflow commonly chains multiple specialized components:

1. An object detector to locate the distress.

2. A segmenter to delineate crack boundaries.

3. A classification component to identify distress type, such as longitudinal, transverse, or alligator cracking.

4. A rule-based or mathematical component to estimate severity and engineering indices.

5. A reporting component to generate records for engineers and asset-management systems.

Relevant conflicts of interest/financial disclosures: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

This modular architecture can deliver high task-specific accuracy, but it also increases labeling effort, software integration cost, maintenance burden, and the risk of downstream error propagation. This review examines the emerging alternative of multimodal Vision Large Language Models (VLLMs), which combine visual perception, semantic reasoning, and structured report generation within a unified inference pathway. It also proposes the Emmaus Pipeline, a fleet-based collection and offline-analysis framework intended to extend monitoring to residential and last-mile roads.

The review is guided by four questions: (1) which elements of a dedicated crack-detection pipeline can be unified by a VLLM; (2) which engineering functions still require calibrated or task-specific components; (3) how can routine public-service fleets support scalable acquisition and localization; and (4) what evaluation and governance controls are required before operational use?

2. MATERIALS AND METHODS

2.1 Review Scope and Evidence Organization

This article is a narrative and architectural review of the nineteen sources listed in the References. The evidence covers pavement-distress detection, road-condition sensing, generative augmentation, distress standards, PCI assessment, segmentation, UAV inspection, lightweight models, multisensor fusion, and concrete-pavement applications [1–19]. The sources were organized according to their function in the monitoring workflow: data acquisition, image enhancement, defect localization, segmentation, classification, severity assessment, geolocation, and reporting.

The review does not estimate a pooled effect size. Instead, it synthesizes reported capabilities and deployment constraints to identify where a unified multimodal model may simplify a conventional system and where dedicated engineering components remain necessary.

2.2 Comparative Evaluation Framework

Traditional deep-learning and VLLM approaches are compared across six technical dimensions: model architecture, labeling requirements, contextual reasoning, output format, edge feasibility, and robustness to illumination. The proposed Emmaus Pipeline is then assessed as a deployment concept, with particular attention to fleet-based acquisition, offline batch processing, geolocation, sensor fusion, and compatibility with asset-management databases. Standard computer-vision metrics and the PCI formulation are retained as the basis for future empirical validation.

3. RESULTS AND DISCUSSION

3.1 The Dedicated Deep-Learning Paradigm

Deep-learning models dominate automated pavement-crack detection because they learn hierarchical visual features that are more robust than handcrafted filters. The conventional architecture nevertheless remains a multi-stage pipeline in which detection, segmentation, classification, severity estimation, and report generation are performed by separate models or rules.

3.1.1 Bounding-Box Detection and Segmentation

Object-detection networks in the YOLO family are widely used for real-time road applications because they combine relatively high inference speed with strong localization performance. Nafaa et al. reported mean average precision at an IoU threshold of 0.50 (mAP50) of up to 96.9% for pavement-crack detection and classification under varied lighting [16].

Where physical width, length, or surface area must be quantified, semantic segmentation is usually preferable. DeepLabv3+ uses atrous spatial pyramid pooling to capture multi-scale context, and contrast-limited adaptive histogram equalization can improve the visibility of thin, low-contrast cracks [9]. U-Net remains an influential encoder-decoder reference architecture for dense segmentation [15].

Specialized architectures further address crack continuity and edge quality. CPCDNet uses crack-alignment and weighted edge-loss mechanisms to preserve narrow structures in textured pavement scenes [11]. Lightweight multi-scale models such as MDCCM reduce parameter count and improve feasibility for constrained edge hardware [13]. UAV-based inspection frameworks extend the same detection and evaluation logic to aerial acquisition [12].

3.1.2 Generative Data Augmentation

Supervised models face a data-imbalance problem because crack pixels occupy only a small portion of most pavement images. Generative adversarial networks and unpaired image-to-image translation can create synthetic distress under different pavement materials, moisture conditions, shadows, and illumination. These examples can regularize training and improve generalization when real pixel-level annotations are scarce [4–6].

3.1.3 Limitations of the Multi-Stage Paradigm

Despite strong benchmark performance, multi-stage systems exhibit several operational limitations:

- Pipeline fragility: an error in detection can prevent segmentation, quantification, and reporting from occurring.
- Labeling overhead: segmentation requires expensive pixel-level masks, and retraining may be needed when pavement materials or imaging conditions change.
- Limited contextual reasoning: conventional outputs are bounding boxes, masks, or class scores and do not directly explain structural implications.
- Integration burden: each component has separate data formats, versioning, monitoring, and failure modes.
- Domain shift: illumination, weather, road markings, and regional pavement textures can reduce transferability.

3.1.4 Data, Annotation, and Domain Shift

The performance of a dedicated model is closely tied to the definition and quality of its labels. Bounding-box datasets are faster to create than pixel masks, but a box includes background pavement and therefore cannot directly represent crack width, branching, or total distressed area. Pixel-level annotation is more informative for engineering measurement, yet thin cracks are difficult to label consistently because the apparent boundary changes with image resolution, blur, shadow, and annotator judgment. A dataset can consequently contain disagreement even before a model is trained.

Class definitions introduce a second source of variation. Longitudinal, transverse, block, and alligator cracking describe morphology, but severity may also depend on width, spalling, interconnectedness, and location within a lane. If a training set collapses these dimensions into a single label, the model can learn a visual class without learning the information required for PCI calculation. Conversely, a highly detailed label taxonomy can create small classes and unstable estimates. The annotation design should therefore begin with the intended maintenance decision rather than with the available model architecture.

Data imbalance occurs at several levels. Most pixels are sound pavement, many captured frames contain no actionable distress, and severe defects are less common than minor ones. A detector trained on a highly curated crack dataset may appear accurate while producing too many false positives when deployed on continuous route video. Negative examples should include sealed cracks, joints, lane markings, tar strips, shadows, leaves, water, utility covers, and surface texture changes. Generative augmentation can increase variation, but synthetic images must be checked to ensure that they preserve physically plausible crack morphology [4–6].

Domain shift is especially important for municipal deployment. Camera height, viewing angle, vehicle speed, compression, season, pavement aggregate, weather, and maintenance practice all change the input distribution. A model trained on one region may detect large cracks in another region while missing hairline distress or confusing repair seams with defects. The review literature emphasizes generalizability as a continuing challenge rather than a solved property of deep learning [10]. Validation should therefore be stratified by road type, lighting, weather, camera configuration, and distress class, with a separate route-based test set that has not contributed frames to training.

Temporal leakage must also be controlled. Consecutive video frames are visually similar, so random frame-level splitting can place nearly identical road views in both training and test sets. This inflates apparent performance and does not measure

transfer to an unseen road segment. Route-level, date-level, or geographic splitting provides a more realistic estimate. For longitudinal monitoring, an additional test should measure whether the model distinguishes genuine deterioration from changes in viewpoint or illumination.

3.1.5 Operational Deployment of Dedicated Models

Dedicated models are attractive for edge deployment because compact detectors can process frames with predictable latency and fixed memory use. The output space is constrained, making failures easier to quantify. A detector can be versioned, tested against a fixed dataset, and configured to reject frames below a confidence threshold. Lightweight segmentation models further reduce resource demand [13]. These properties are valuable where a vehicle must alert an operator in real time or where connectivity is limited.

The operational challenge is the number of dependent services. A frame-quality filter, detector, segmenter, classifier, dimension estimator, geolocation component, database adapter, and reporting service may each have different update cycles. An improvement in one component can change the input distribution of the next. The system therefore requires end-to-end regression tests, not only model-level metrics. Monitoring should record the proportion of frames reaching each stage, the reasons for rejection, and the distribution of confidence scores by route and environmental condition.

A further distinction is required between visual detection and engineering diagnosis. A CNN may reliably identify a crack pattern without establishing the underlying structural cause. Likewise, a high-confidence mask does not automatically produce a valid maintenance recommendation. Dedicated models are most defensible when their output is limited to a clearly validated visual task and when engineering rules or human review convert that output into action. This separation of responsibilities provides a useful benchmark for judging whether a VLLM genuinely simplifies the pipeline or merely hides multiple uncertain steps inside one model.

3.2 The Shift to Multimodal Vision Large Language Models

A multimodal VLLM typically combines a visual encoder with an autoregressive language model through a projection or cross-attention mechanism. The model can interpret an image and an engineering prompt jointly, then generate natural-language and structured outputs. Figure 1 contrasts this unified pathway with a conventional chain of specialized models.

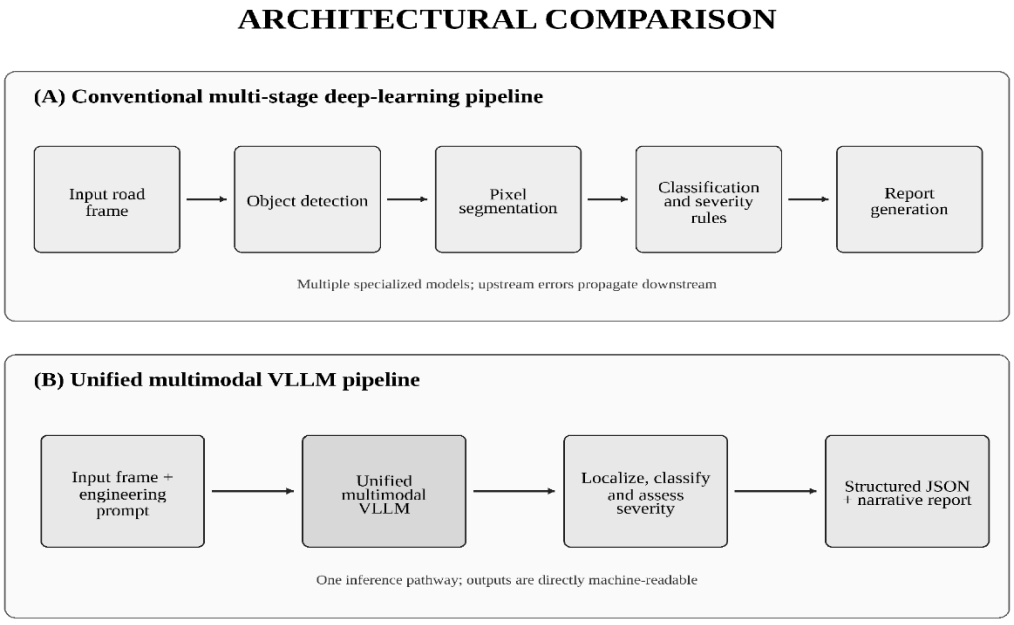

Figure 1. Architectural comparison between a conventional multi-stage deep-learning pipeline and a unified VLLM pipeline.

3.2.1 Unification of the Inference Pipeline

A single VLLM can be prompted to perform several tasks within one response:

1. Visual localization: identify a crack region using normalized bounding-box or point coordinates.

2. Classification and semantic description: describe crack morphology and orientation.

3. Severity reasoning: assign a preliminary severity category using visible width, extent, and contextual cues.

4. Structured output: return a predefined JSON schema for direct ingestion into an asset-management database.

This unification reduces the number of separately maintained inference services and makes the output more accessible to engineers. It does not, however, guarantee metrically accurate crack dimensions. Physical measurement still requires calibrated imaging, scale references, or geometric reconstruction.

3.2.2 In-Context Learning and Generalizability

Unlike a fixed CNN classifier, a VLLM can adapt its output behavior through system instructions and a small number of in-context examples. This is potentially useful when moving between asphalt, concrete slabs, culverts, and region-specific surface textures [18,19]. Nevertheless, zero-shot performance should be treated as a hypothesis to be validated rather than as a substitute for local testing. Prompt sensitivity, model updates, and non-deterministic generation can affect reproducibility.

3.2.3 Capabilities, Uncertainty, and Failure Modes

The central advantage of a VLLM is task composability. The same input can be used to request a defect label, an explanation, a structured record, a comparison with an earlier observation, or a question-answering response for an engineer. This flexibility can reduce the amount of custom code required to connect perception with reporting. It is particularly useful for long-tail cases in which an image contains several interacting features, such as a crack adjacent to a patch, standing water, and a lane marking.

However, a fluent response is not equivalent to a calibrated measurement. A VLLM may describe a crack as wide or severe without a known scale, or it may infer a causal mechanism that is not visible in the frame. The model can also omit a small defect when a more visually salient object dominates the scene. These failures differ from a conventional detector because they may appear as plausible prose rather than as an obvious missing box. Evaluation must therefore score every required output field and not rely on the apparent quality of the narrative.

Spatial grounding is another limitation. Some multimodal models can return bounding boxes or point coordinates, but coordinate precision may be lower than that of a detector trained specifically for localization. Small coordinate errors are important when the target is a narrow crack. A coarse box may be sufficient for triage but inadequate for width or density estimation. A practical architecture can use the VLLM to identify the relevant distress and a dedicated segmentation component to refine geometry. The resulting system remains simpler than a fully independent chain if the VLLM replaces classification, explanation, and report generation.

Uncertainty should be represented explicitly. A single confidence score is insufficient because the model can be certain about the presence of a defect but uncertain about its type or severity. The output schema should separate detection confidence, classification confidence, severity confidence, and localization quality. It should also include a reason for uncertainty, such as blur, obstruction, poor illumination, lack of scale, or conflicting visual cues. Frames below a defined quality level should be routed to a human or to a dedicated model rather than forced into a complete report.

Reproducibility requires control over the model version, prompt, decoding configuration, image preprocessing, and output parser. A service provider may update a hosted model without changing the municipal application code. The same image can then produce a different result. For research comparisons, the full inference configuration should be archived and a stable test set should be reprocessed after each model change. Where deterministic behavior is essential, a locally hosted model or a dedicated classifier may remain preferable.

The language component also creates a risk of unsupported causal interpretation. Statements about base-layer failure, drainage problems, or axle loading may be useful hypotheses, but they should not be stored as confirmed diagnoses unless corroborated by engineering evidence. A controlled vocabulary can distinguish observed features from inferred causes and recommended inspections. This separation allows the system to retain the explanatory benefit of a VLLM without overstating what can be concluded from a single image.

3.2.4 Prompt Design and Structured Output Validation

A production prompt should define the role of the model, the permitted distress taxonomy, the severity rules, the coordinate convention, the required output schema, and the conditions under which the model must abstain. The instruction should require concise evidence for each classification and prohibit unsupported physical dimensions. Few-shot examples should include both positive and negative cases, including sealed cracks, shadows, joints, and uncertain images.

Structured output reduces ambiguity only when it is validated. The parser should reject malformed JSON, missing mandatory fields, out-of-range coordinates, invalid class labels, and mutually inconsistent values. A record that claims no defect but includes a high-severity crack should fail validation. The application can automatically retry with a correction prompt, but repeated failure should create an exception for review rather than silently converting free text into a database value.

Table 2 proposes a minimum schema for the Emmaus Pipeline. The exact field names may vary, but the distinction between observation, uncertainty, and engineering action should be preserved.

Field	Purpose
Frame and route identifiers	Links the result to the source image, vehicle, route, and capture time.
Image-quality status	Records blur, obstruction, illumination, and whether the model should abstain.
Distress observation	Stores presence, type, orientation, and visible morphology using a controlled vocabulary.
Spatial grounding	Stores normalized box or point coordinates and the coordinate convention.
Severity and evidence	Separates the preliminary severity category from the visible evidence used to assign it.
Uncertainty fields	Records confidence by task and the reason for uncertainty.
Geolocation fields	Stores GPS estimate, interpolation method, and estimated spatial error.
Engineering disposition	Routes the record to monitor, review, inspect, or prioritize without asserting an unsupported diagnosis.
Model provenance	Stores model version, prompt version, and processing timestamp for auditability.

Table 2. Recommended Structured Output Fields For VLLM Analysis

A validated schema also supports comparative experiments. The same route frame can be processed by different models, and field-level accuracy can be calculated for crack presence, type, severity, and location. Narrative quality may be assessed separately by engineers, but it should not replace objective scoring of the structured fields.

3.3 The Proposed Emmaus Pipeline

The Emmaus Pipeline is a decentralized road-monitoring concept that separates inexpensive collection from computationally intensive analysis. Existing public-service vehicles collect synchronized road video and location data during routine operations; processing is performed after the vehicles return to a depot. Figure 2 presents the proposed flow.

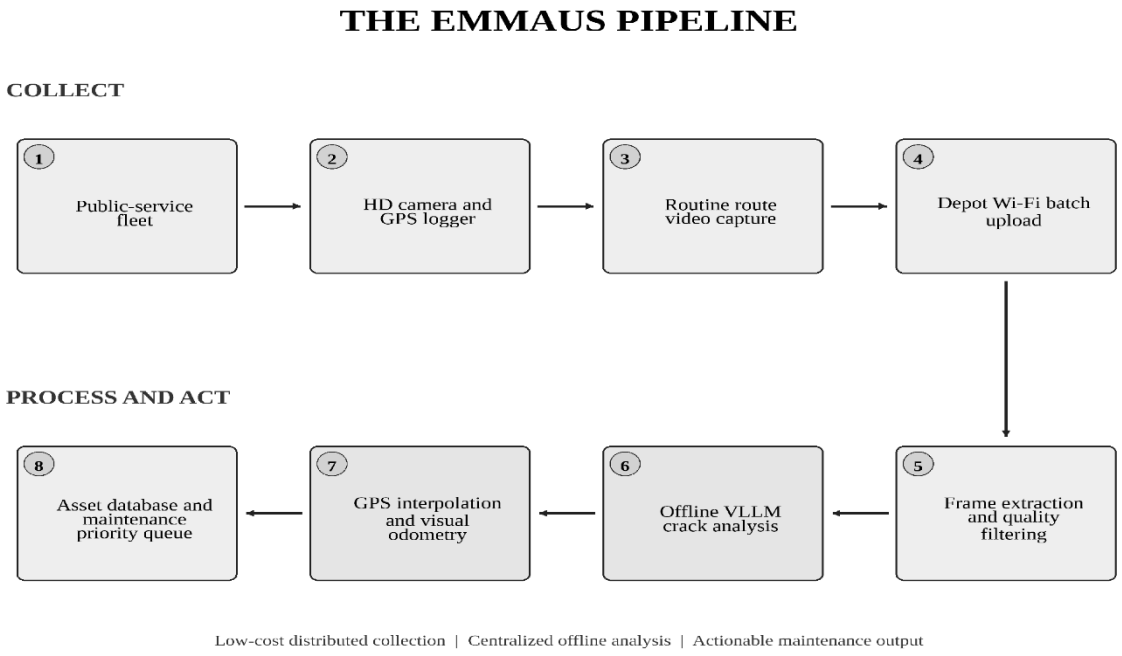

Figure 2. Flow diagram of the proposed Emmaus Pipeline from fleet-based collection to maintenance prioritization.

3.3.1 Fleet-Based Decentralized Data Collection

Instead of relying exclusively on specialized pavement-inspection vehicles, the Emmaus Pipeline uses routine municipal or public-service fleets, including garbage trucks, postal vans, and highway-patrol vehicles. Other mobile platforms, including coordinated robots and UAVs, illustrate the broader trend toward distributed infrastructure sensing [3,12].

Consumer-grade high-definition cameras and GPS loggers can passively record road surfaces while these vehicles traverse local streets. This approach may improve coverage of residential roads that are inspected less frequently, while avoiding the power and thermal constraints of continuous onboard inference.

3.3.2 Offline Batch Processing

Video and synchronized GPS logs are uploaded over depot Wi-Fi at the end of a shift or on a scheduled batch cycle. The processing system samples frames at distance- or time-based intervals, removes frames affected by severe motion blur or obstruction, and submits the retained frames to an offline VLLM.

The output should follow a controlled schema containing the defect type, location estimate, confidence, severity category, textual description, and links to the source frame. The schema enables quality assurance, database ingestion, and later comparison with dedicated detector outputs.

3.3.3 Resolving the Localization Challenge

The principal technical hurdle is precise spatial localization. Standard consumer GPS commonly has meter-level uncertainty, and signal quality may deteriorate in urban canyons or under dense tree cover. The Emmaus Pipeline therefore uses a layered strategy:

1. Timestamp-GPS interpolation: match each video-frame timestamp to the vehicle trajectory and interpolate between adjacent GPS observations.

2. Visual odometry and structure from motion: estimate relative vehicle displacement between GPS updates and smooth short-term drift.

3. Multisensor fusion: combine video with inertial, acoustic, vibration, laser-profile, or ground-penetrating-radar signals where higher fidelity is required [14,17].

4. Repeated-route reconciliation: compare observations from multiple passes and retain the most spatially consistent location estimate.

3.3.4 Data Lifecycle and Municipal Integration

The Emmaus Pipeline depends on a disciplined data lifecycle. Capture begins with a route identifier, vehicle identifier, calibrated clock, camera configuration, and GPS logging status. At upload, the system verifies file completeness and synchronizes timestamps before frame extraction. Each retained frame receives a stable identifier so that model outputs, human corrections, and later inspections can be traced to the original evidence.

Frame sampling should reflect the intended use. A fixed time interval produces more samples when a vehicle is moving slowly, whereas a fixed distance interval provides more uniform road coverage. A hybrid rule can retain frames after a specified distance while adding samples near a detected defect. The system should avoid discarding the surrounding sequence because adjacent frames can improve visual odometry and allow a reviewer to distinguish a persistent crack from a transient obstruction.

The asset-management database should not store only the latest label. It should maintain an observation history for each road segment, including capture conditions, model version, review status, and links to prior records. A new observation can then be compared with the previous condition to identify persistence or apparent progression. When location uncertainty causes two records to overlap, the system should preserve both observations until a reconciliation rule or human review confirms that they represent the same defect.

Integration with maintenance planning requires a distinction between condition evidence and work priority. A high-severity observation on a low-volume road may not receive the same priority as a moderate defect at a critical intersection. The VLLM output should therefore feed an existing prioritization process rather than replace it. Budget, traffic exposure, safety consequence, drainage, planned resurfacing, and inspection confidence remain separate decision variables.

Batch processing also enables quality control at municipal scale. Daily or weekly dashboards can report route coverage, upload failures, frame-rejection rates, defect counts, spatial uncertainty, and the proportion of records awaiting review. These operational measures reveal whether the collection system is functioning before model accuracy is considered. A technically strong model cannot compensate for missing routes, unsynchronized clocks, obstructed cameras, or incomplete uploads.

3.3.5 Privacy, Security, and Human Oversight

Road video can incidentally capture people, vehicle plates, house fronts, and location patterns. Privacy protection should therefore begin before analysis. Municipal policy can define capture angles, retention periods, access roles, encryption, and automatic redaction of identifiable information. Raw video may be retained only long enough to support quality assurance, while cropped pavement frames and structured results can be stored for longer-term condition analysis.

Security controls are also required because a manipulated route file or model output could distort maintenance priorities. Uploads should be authenticated, records should be immutable after approval, and every automated change should be logged. Model prompts and schemas are part of the controlled system configuration and should be

versioned like software. Human reviewers should be able to see the source frame, uncertainty fields, and prior observations before accepting a high-priority recommendation.

Human oversight should be risk-based rather than universal. Low-confidence records, severe defects, unusual classes, large changes from prior observations, and geolocation conflicts can be routed for review. High-confidence low-severity observations can be sampled for audit. This approach preserves scalability while creating evidence about real-world error patterns that can guide model improvement.

3.4 Quantitative Framework for Pavement Assessment

3.4.1 Pavement Condition Index

The Pavement Condition Index (PCI) is a numerical indicator from 0 to 100, where 100 represents a pavement in excellent condition and 0 represents failure. PCI is derived from distress type, severity, density, and empirically determined deduct values [7,8]. A simplified expression is:

$$PCI = 100 - \Sigma^{p}_{i=1} \Sigma^{q}_{j=1} a(T_i, S_j, D_{ij})$$

where T_i is distress type i; S_j is severity level j; D_{ij} is distress density; and $a(T_i, S_j, D_{ij})$ is the deduct value taken from the applicable standard curve. Distress density may be expressed as:

$$D_{ij} = (\text{total area or length of distress i / total sample-unit area}) \times 100\%$$

A VLLM can supply preliminary type, severity, and region estimates, but automated PCI calculation should use calibrated dimensions and the applicable standard deduct-value procedure. Uncalibrated visual estimates should not be represented as engineering measurements.

3.4.2 Evaluation Metrics

Future experiments should compare VLLM outputs with dedicated deep-learning systems using established detection and segmentation metrics:

$$IoU = |A \cap B| / |A \cup B|$$

$$Precision = TP / (TP + FP) \qquad Recall = TP / (TP + FN)$$

$$F_1 = 2 \times (Precision \times Recall) / (Precision + Recall)$$

Here, A is the ground-truth mask, B is the predicted mask, TP denotes true positives, FP denotes false positives, and FN denotes false negatives. An operational evaluation should also report localization error, severity agreement, schema-valid output rate, inference latency, processing cost, and the proportion of frames rejected for quality reasons.

3.4.3 Proposed Empirical Validation Protocol

A credible evaluation of the Emmaus Pipeline should separate model accuracy from system performance. The first level is frame quality: whether the camera, exposure, motion, and obstruction permit assessment. The second is visual perception: whether distress is detected, localized, and classified correctly. The third is engineering interpretation: whether severity and PCI-related variables agree with a calibrated reference. The fourth is operational performance: whether observations are geolocated, stored, reviewed, and converted into useful maintenance actions.

The test dataset should be collected from complete routes rather than only from selected defect images. This preserves the true prevalence of distress and measures the false-positive burden on routine video. Routes should include different pavement materials, road classes, weather conditions, times of day, camera positions, and vehicle speeds. Ground truth should be established by trained annotators using written definitions, with engineering review for ambiguous distress and physical measurement for a subset of sites.

The comparison should include at least three configurations: a dedicated detector or segmentation pipeline, a VLLM-only pipeline, and a hybrid pipeline in which the VLLM is combined with a task-specific localization or measurement model. All configurations should receive the same frames and use the same route-level data split. The evaluation should record both average performance and failure concentration by distress type, lighting, road class, and image quality.

Geolocation should be validated independently of visual classification. Surveyed reference points or manually verified map locations can be used to

calculate horizontal position error. The experiment should compare raw GPS, timestamp interpolation, visual-odometry correction, and repeated-route reconciliation. A system may detect the correct crack but still fail operationally if the resulting work order points to the wrong road segment.

Severity agreement should be assessed with a confusion matrix and an ordinal measure because confusing low with medium severity is less serious than confusing low with high. For PCI-related use, estimated distress dimensions should be compared with calibrated field measurements. The study should clearly distinguish image-level classification from sample-unit PCI calculation; the latter requires aggregation, density calculation, and standard deduct-value procedures [7,8].

Reliability testing should repeat inference on the same frames and, where possible, across model versions. The rate of schema-valid responses, abstentions, inconsistent labels, and unsupported explanations should be reported. Cost analysis should include storage, transfer, compute, review time, and the operational cost of false positives and missed defects. These measures permit a fair comparison because a lower model error does not necessarily produce a lower total monitoring cost.

Table 3 summarizes the recommended validation layers and primary measures.

Validation layer	Primary reference	Recommended measures
Capture quality	Human frame-quality labels	Usable-frame rate; blur and obstruction rejection; route coverage.
Detection and classification	Expert image annotation	Precision; recall; F1-score; class confusion; false positives per route kilometre.
Segmentation and geometry	Pixel masks and calibrated measurements	IoU; boundary error; width and length error; distress-area error.
Severity and PCI variables	Engineer rating and field survey	Ordinal agreement; severity confusion; density error; PCI difference.
Geolocation	Surveyed or manually verified locations	Median and 95th-percentile position error; road-segment assignment accuracy.
Structured output	Schema and business rules	Valid-record rate; missing fields; contradictory fields; abstention rate.
Operational performance	System logs and maintenance workflow	Latency; cost per kilometre; review workload; time from capture to action.
Reproducibility and governance	Repeated runs and version records	Run-to-run agreement; version drift; audit completeness; privacy exceptions.

Table 3. Proposed Validation Layers For The Emmaus Pipeline

A field trial can be staged. An initial pilot can test a limited number of routes and focus on data synchronization, schema validity, and reviewer workflow. A second phase can expand to diverse conditions and compare model configurations. Only after stable performance should the system influence maintenance prioritization. This progression reduces the risk of interpreting a successful image demonstration as evidence of a reliable municipal system.

3.5 Comparative Synthesis

Table 1 summarizes the principal trade-offs. The comparison indicates that VLLMs can simplify orchestration and reporting, whereas dedicated models retain advantages in deterministic execution, calibrated pixel-level measurement, and efficient edge deployment.

Technical dimension	Traditional deep learning (YOLO / DeepLab)	Multimodal VLLM (Emmaus framework)
Model architecture	Multi-stage: detect, segment, classify, quantify, and report.	Unified vision-language inference with structured output.
Labeling requirements	High for bounding boxes and especially pixel-level masks.	Lower for initial prototyping; few-shot prompts may guide behavior, but validation data remain necessary.
Contextual reasoning	Limited to trained outputs and programmed rules.	Can generate semantic descriptions and contextual explanations.
Output format	Boxes, masks, labels, and numerical arrays.	Coordinates, labels, JSON records, and narrative reports.
Edge feasibility	Strong for lightweight and quantized models.	Usually better suited to cloud or local-server batch processing.
Illumination robustness	Often requires augmentation or preprocessing.	Broad pretraining may improve tolerance, but performance must be measured locally.

Table 1. Technical Comparison Of Traditional Deep Learning And The Emmaus VLLM Framework

The synthesis suggests that the most defensible near-term architecture may be hybrid. A dedicated detector or segmenter can provide reproducible geometry, while a VLLM can interpret context, validate classifications, explain uncertainty, and produce structured reports. The Emmaus Pipeline is therefore best viewed as a testable architecture rather than as evidence that dedicated models are no longer required.

Architecture should be selected according to the decision being supported. For rapid network screening, a VLLM may provide sufficient localization and descriptive triage. For crack-width measurement, density calculation, or contractual acceptance, calibrated segmentation and field verification are more appropriate. The same municipal system can use different processing paths based on image quality, distress severity, and required measurement precision.

The principal simplification offered by a VLLM is not necessarily the removal of every dedicated model. It is the consolidation of semantic tasks that otherwise require separate classifiers, rules, report templates, and user interfaces. Even when a segmenter remains, a VLLM can convert the mask, metadata, and prior observation into a consistent engineering record. This narrower claim is more testable and avoids assuming

that general-purpose visual reasoning already matches specialist measurement models.

Operational value also depends on coverage and timeliness. A modestly accurate system that repeatedly observes the full local network may identify deterioration earlier than a highly accurate system deployed infrequently on selected roads. The Emmaus concept emphasizes this coverage advantage. Its research contribution should therefore be evaluated at both the frame level and the network level, including kilometres observed, repeat frequency, time to review, and the number of actionable defects confirmed.

Deployment decisions should also consider data governance. Road video can capture faces, license plates, private property, and location histories. Municipal implementations require retention limits, access controls, redaction procedures, model-version records, audit trails, and human review of maintenance priorities.

3.6 Limitations of This Review

This article is a narrative review and conceptual synthesis rather than a systematic review or meta-analysis. The included studies use different datasets, distress definitions, camera configurations, train-test splits, and performance measures, so reported accuracy values are not directly comparable. Several references address concrete cracking, mobile sensing, or related road-monitoring tasks rather than the complete Emmaus workflow. Their relevance lies in the technical component they inform.

The proposed VLLM capabilities have not been validated through a common field experiment in this article. Claims about simplification therefore describe architectural potential, not demonstrated superiority. Commercial and open-source multimodal models also evolve rapidly, which can change cost, grounding capability, context length, and deployment feasibility. The validation protocol is intended to make future comparisons reproducible despite this change.

Finally, the review focuses on visible surface distress. Subsurface defects, structural capacity, drainage, skid resistance, rut depth, and other condition variables may require sensors or inspections beyond ordinary road video. The Emmaus Pipeline should be integrated with, not substituted for, established engineering assessment where those variables determine maintenance decisions.

CONCLUSION

The transition from dedicated deep-learning architectures to multimodal VLLMs creates a credible opportunity to simplify pavement-monitoring workflows. A unified model can combine visual interpretation, semantic description, preliminary severity reasoning, and machine-readable reporting. When paired with routine public-service fleets and depot-based processing, this capability could lower the operational barrier to broader local-road coverage.

The review also identifies important constraints. VLLMs require substantial compute, may produce inconsistent outputs, and cannot infer reliable physical dimensions from uncalibrated imagery. Engineering validation, controlled output schemas, model-version governance, and human oversight are therefore essential. Future research should prioritize:

1. Multisensor temporal synchronization of video, GPS, IMU, acoustic, and pavement-response signals.

2. Distillation and quantization of multimodal models for lower-cost local or edge processing.

3. Longitudinal change analysis that compares repeated observations of the same road segment.

4. Field benchmarking against dedicated detectors, calibrated measurements, and PCI surveys.

5. Governance protocols for privacy, traceability, uncertainty reporting, and maintenance decision review.

The Emmaus Pipeline provides a practical framework for such evaluation by linking distributed data collection, offline multimodal analysis, geolocation, and asset-management prioritization within one end-to-end research design.

Acknowledgement

The author acknowledges the researchers and public agencies whose published work forms the basis of this review.

Conflicts of Interest

The author declares no conflict of interest related to this article.

Ethical Statement

Ethical approval was not required because this article reviews published literature and proposes a conceptual system architecture; it reports no experiments involving human participants or animals.

REFERENCES

1. Y. Du, N. Pan, Z. Xu, F. Deng, and Y. Shen, "Pavement distress detection and classification based on YOLO network," International Journal of Pavement Engineering, vol. 22, no. 13, pp. 1659–1672, 2021. https://doi.org/10.1080/10298436.2020.1714047
2. E. Ranyal, A. Sadhu, and K. Jain, "Road Condition Monitoring Using Smart Sensing and Artificial Intelligence: A Review," Sensors, vol. 22, no. 8, Art. no. 3044, 2022. https://doi.org/10.3390/s22083044
3. M. Alkhedher, A. Alsit, M. Alhalabi, S. AlKheder, A. Gad, and M. Ghazal, "Novel pavement crack detection sensor using coordinated mobile robots," Transportation Research Part C: Emerging Technologies, vol. 172, Art. no. 105021, 2025. https://doi.org/10.1016/j.trc.2025.105021
4. H. Maeda, Y. Sekimoto, T. Seto, T. Kashiyama, and H. Omata, "Generative adversarial network for road damage detection," Computer-Aided Civil and Infrastructure Engineering, vol. 36, no. 1, pp. 47–60, 2021. https://doi.org/10.1111/mice.12561
5. T. W. Muturi, "Rigid Pavement Crack Detection Utilizing Generative Adversarial Networks," Master's Thesis, Middle East Technical University, Department of Civil Engineering, Jan. 2023.
6. J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), Oct. 2017, pp. 2223–2232. https://doi.org/10.1109/ICCV.2017.244
7. J. S. Miller and W. Y. Bellinger, Distress Identification Manual for the Long-Term Pavement Performance Program, 5th revised ed., FHWA-HRT-13-092, Federal Highway Administration, U.S. Department of Transportation, Jun. 2014.
8. S. M. Piryonesi and T. E. El-Diraby, "Data Analytics in Asset Management: Cost-Effective Prediction of the Pavement Condition Index," Journal of Infrastructure Systems, vol. 26, no. 1, Art. no. 04019036, 2020. https://doi.org/10.1061/(ASCE)IS.1943-555X.0000512
9. X. Wang, T. Wang, and J. Li, "Advanced crack detection and quantification strategy based on CLAHE enhanced DeepLabv3+," Engineering Applications of Artificial Intelligence, vol. 126, Part B, Art. no. 106880, 2023. https://doi.org/10.1016/j.engappai.2023.106880
10. X. Zhang, H. Wang, Y.-A. Hsieh, Z. Yang, A. Yezzi, and Y.-C. Tsai, "Deep Learning for Crack Detection: A Review of Learning Paradigms, Generalizability, and Datasets," arXiv Preprint, arXiv:2508.10256, 2025. https://doi.org/10.48550/arXiv.2508.10256
11. J. Zhang, S. Sun, W. Song, Y. Li, and Q. Teng, "A novel convolutional neural network for enhancing the continuity of pavement crack detection," Scientific Reports, vol. 14, Art. no. 211, 2024. https://doi.org/10.1038/s41598-024-81119-1
12. X. Chen, C. Liu, L. Chen, X. Zhu, Y. Zhang, and C. Wang, "A Pavement Crack Detection and Evaluation Framework for a UAV Inspection System Based on Deep Learning," Applied Sciences, vol. 14, no. 3, Art. no. 1157, 2024. https://doi.org/10.3390/app14031157
13. Z. Gu, T. Li, Q. Xiao, J. Chen, G. Ding, and H. Ding, "MDCCM: a lightweight multi-scale model for high-accuracy pavement crack detection," Signal, Image and Video Processing, vol. 19, Art. no. 488, 2025. https://doi.org/10.1007/s11760-025-04064-0
14. Z. Wang, D. Qiu, R. Wu, Y. Shi, and W. Niu, "Research on Road Crack Detection Based on RGB-LPC-GPR Data Fusion," International Journal of Automated Infrastructure Monitoring, vol. 8, no. 2, pp. 88–101, 2025.

15. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
16. S. Nafaa, H. Essam, H. I. Ashqar, K. Ashour, D. Emad, A. A. Hassan, R. Mohamed, M. Elhenawy, and T. I. Alhadidi, "Automated Pavement Cracks Detection and Classification Using Deep Learning," arXiv Preprint, arXiv:2406.07674, Jun. 2024. https://doi.org/10.48550/arXiv.2406.07674
17. H. I. Ashqar et al., "Vulnerable road user detection using smartphone sensors and recurrence quantification analysis," in 2019 IEEE Intelligent Transportation Systems Conference (ITSC), Oct. 2019, pp. 1054–1059. https://doi.org/10.1109/ITSC.2019.8917227
18. S. Biswas and P. Sengupta, "CNN-Based Crack Detection of Reinforced Concrete Slab Culverts," in Recent Developments in Structural Engineering, Volume 1 (SEC 2023), Lecture Notes in Civil Engineering, vol. 52, Springer, Singapore, 2024, pp. 623–631. https://doi.org/10.1007/978-981-99-9625-4_59
19. E. Ranyal, V. Ranyal, and K. Jain, "AI Based Non-contact Crack Detection and Measurement in Concrete Pavements," in Proceedings of the Canadian Society for Civil Engineering Annual Conference 2023, Volume 12 (CSCE 2023), Lecture Notes in Civil Engineering, vol. 506, Springer, Cham, 2025, pp. 141–154. https://doi.org/10.1007/978-3-031-61535-1_12

HOW TO CITE: Alfred Patrick Patric*, Simplifying Pavement Crack Detection: From Deep Learning Pipelines To Multimodal Vision-Language Models, Int. J. Sci. R. Tech., 2026, 3 (7), 672-685. https://doi.org/10.5281/zenodo.21471379