

Reviewer 2

We sincerely thank Reviewer 2 for the thorough, technically detailed, and critical assessment of our manuscript. Although the comments raised significant concerns regarding the positioning, validation scope, and deployability of the proposed framework, they proved extremely valuable in helping us identify conceptual ambiguities and structural weaknesses in the original submission.

Addressing these comments required a substantial rethinking of how the geospatial component, instance-level evaluation, and end-to-end pipeline demonstration were framed and validated. As a result, the manuscript has been carefully revised and strengthened, with clearer separation between deployable inference stages and evaluation-only procedures, a more precise positioning of the GIS integration as illustrative, and improved internal consistency across the experimental analysis.

We believe that these revisions have significantly improved the technical clarity, honesty, and practical relevance of the work.

R2.1 Comment: “The current work does not provide evidence strong enough for publication, mainly because the geospatial part is not validated in a way that matches the paper’s own positioning, and because the end-to-end demonstration is built on evaluation steps that depend on ground truth in a way that cannot be reproduced at deployment time. The technical problem is the geospatial mapping setup . . . The work therefore demonstrates a coordinate conversion procedure, not localization accuracy in a real UAV inspection setting.”

We thank the reviewer for this important and closely related comment. It helped us identify and clarify a key ambiguity in the original manuscript concerning the role, scope, and validation of the geospatial component.

We fully agree that, in the absence of real UAV navigation metadata, camera calibration, flight altitude, and pose information, it is not possible to validate absolute geospatial localisation accuracy in a real deployment scenario. As correctly noted by the reviewer, the use of synthetic waypoints derived from a road model together with an assumed ground sampling distance demonstrates a coordinate conversion and GIS integration procedure, rather than sensor-accurate localisation.

In the revised manuscript, we have therefore substantially revised both the positioning and structure of the geospatial component to ensure that it accurately reflects the scope of the validation performed in this work. In particular, we now: (i) explicitly clarify that the core technical contribution and quantitative validation focus on the perception and geometric processing stages (semantic segmentation, instance-level analysis, contour extraction, and polygonal vector representation); (ii) reposition the geospatial mapping stage as an illustrative integration step rather than

a localisation accuracy experiment; and (iii) separate GIS-related results into a dedicated illustrative section.

Excerpt from the revised manuscript (Section 1, Page 3, Paragraph 11, Lines 97–103):

“The primary focus lies on the image-based stages of the pipeline, namely semantic segmentation, contour extraction, and the geospatial representation of pavement defects. The geospatial component is introduced to demonstrate how the resulting vectorised defects can be integrated into standard GIS-based inspection workflows, supporting semantic-level information exchange and practical use in scenarios such as routine condition monitoring, preventive maintenance planning, and post-event damage assessment.”

Additional changes:

- Section Abstract, Page 1, Lines 14–16: explicit clarification of GIS integration as an illustrative example of the proposed pipeline.
- New Section 5, *“Full Pipeline Execution: An Illustrative Example”*: GIS results are presented as an end-to-end integration demonstration rather than a localisation validation.

In addition, this limitation is now explicitly acknowledged and discussed in the manuscript, both in the Discussion and in the Conclusions, to clearly delimit the scope of the geospatial contribution. Section 6 (Discussion) analyses the implications of relying on synthetic georeferencing for result interpretation, while Section 7 (Conclusions) summarises this limitation and identifies the integration of real UAV navigation data as a key direction for future work.

Specifically, Section 6 (Discussion), Page 32, Paragraph 12, Lines 943–945 discusses the illustrative nature of the geospatial mapping, and Section 7 (Conclusions), Page 34, Paragraph 5, Lines 1009–1011 summarises this limitation and outlines future extensions.

R2.2 Comment: “A second problem is that the GIS output is produced only after filtering predictions using a ground truth dependent instance validation step... This makes the final GIS layer look clean, but it is not a deployable pipeline because ground truth is not available at inference time. It also hides false positives.”

We thank the reviewer for identifying this limitation and ambiguity in the original manuscript.

In the revised version, we have explicitly decoupled instance-level evaluation from the deployable inference pipeline. No ground-truth-dependent filtering is applied to the GIS outputs.

Excerpt from the revised manuscript (Section 3.4.2, Page 11, Paragraph 8, Lines 362–368):

“All subsequent stages of the pipeline, including contour extraction, polygon generation, and geospatial mapping, operate on the complete set of predicted defect instances without ground-truth-dependent filtering. This separation ensures that the evaluation procedure does not influence the inference pipeline and that the resulting GIS outputs faithfully reflect a realistic deployment scenario in which ground truth is unavailable.”

This separation between evaluation and inference is already established at the instance-level evaluation stage. As clarified in the revised methodology, instance matching is performed exclusively for analysis purposes and does not influence any downstream processing step.

Excerpt from the revised manuscript (Section 3.4.2, Page 10, Paragraph 7, Lines 353–358):

“Instance-level evaluation is used exclusively for evaluation and analysis purposes, and not as a filtering mechanism for the deployable pipeline. The Hungarian-based matching and IoU thresholds are employed to categorise predicted instances as true positives, false positives, or false negatives when computing instance-level performance metrics.”

Additional clarifications:

- Section 3.9 (Contour Extraction and Shapefile Generation), Page 13, Paragraph 6, Lines 473–477: explicitly states that all predicted instances are retained for contour extraction and vectorisation.

“For analysis purposes only, predicted polygons may be annotated with instance-level matching information (...) This information is not used to filter or modify the vector outputs, ensuring that the GIS layer reflects the complete set of predicted defects.”

- Section 5.2 (Illustrative GIS-Based Representation): all predicted geometries are displayed in the GIS layer. Colouring in Fig. 14 is used only for interpretability (IoU ≥ 0.5 vs. < 0.5), explicitly exposing false positives.

R2.3 Comment: “There are also concerns about evaluation rigor and internal consistency... the dataset is small... and some reported counts appear inconsistent with the stated split sizes.”

We thank the reviewer for this insightful comment and fully agree that the original evaluation relied too heavily on a very small test set, which could obscure the stability of the reported metrics.

Motivated directly by this observation, we introduced a dedicated stability analysis with respect to test set size. Rather than relying on a single train-validation-test split, we evaluate multiple configurations with increasing test set proportions (5%, 15%, 20%, and 25%), while keeping the training and validation protocol unchanged. This design allows us to explicitly assess the sensitivity of segmentation metrics and the selected operating threshold to the size of the evaluation set.

In addition, the dataset description has been revised to clearly state the final split configuration used for reporting results. The dataset is split into training, validation, and test subsets following a 65/15/20 ratio (260/60/80), which is now used consistently throughout the manuscript and forms the basis for the final quantitative evaluation.

Excerpt from the revised manuscript (Section 3.6, Page 12, Paragraph 4, Lines 419–420):

“This protocol allows disentangling the effect of test set size from other sources of variability and enables a principled analysis of metric stability as the evaluation set increases.”

Finally, all instance-level statistics were carefully revised to ensure internal consistency with the stated split sizes. Reported counts are now explicitly tied to their corresponding evaluation configuration. For the final test set, 68 out of the 80 images contain at least one true-positive instance, corresponding to 85% of the evaluation set, as now correctly reported in Section 4.4.

R2.4 Comment: “The instance level evaluation design is also not well aligned with crack geometry... the role of morphological operations is unclear... and the term lightweight is not supported quantitatively.”

We thank the reviewer for this detailed and constructive comment.

Instance-level evaluation. We fully agree that axis-aligned bounding-box IoU is a conservative similarity measure for thin, elongated, and fragmented crack-like defects. This limitation is now explicitly acknowledged. In Section 4.4 and in the Discussion (Section 6), we clarify that

pixel-wise semantic segmentation metrics constitute the primary evaluation paradigm, while instance-level bounding-box metrics are used as a complementary and intentionally conservative proxy for localisation behaviour.

Excerpt from the revised manuscript (Section 4.4), Page 24, Paragraph 11, Lines 716–718):

“Overall, these examples highlight that instance-level IoU for crack-like defects is strongly influenced by component fragmentation and bounding-box geometry, and should therefore be interpreted with caution in conjunction with pixel-wise segmentation quality.”

Excerpt from the revised manuscript (Section 6), Page 32, Paragraph 7, Lines 915–920):

“These observations underline that instance-level IoU based on bounding boxes should not be interpreted as a definitive measure of detection quality for crack-like defects. Instead, it provides complementary information on object-level localisation behaviour and should be analysed jointly with pixel-wise metrics and qualitative inspection. This limitation motivates future work on geometry-aware instance similarity measures better suited to linear or high-aspect-ratio structures.”

Morphological post-processing and ablation analysis.

We thank the reviewer for raising this point, which directly motivated a systematic re-evaluation of the post-processing design choices used in the original pipeline.

In the initial version of the framework, lightweight morphological operations were applied prior to instance extraction with the practical goal of stabilising thin crack structures, improving contour continuity, and producing visually clean and interpretable GIS overlays. At that stage, greater emphasis was placed on geometric regularity and qualitative GIS readability than on preserving raw pixel-level performance.

Following the reviewer’s comment, we conducted a dedicated and exhaustive post-processing ablation study to explicitly disentangle the contribution of the network from heuristic refinement (Sections 3.5 and 4.2). This study systematically evaluates four commonly used post-processing variants—raw thresholding, area filtering, morphological operations, and their combination—under a consistent threshold selection protocol based on validation Dice maximisation.

The results reveal that, while morphological operations may slightly regularise contours visually, they consistently degrade pixel-wise segmentation metrics and reduce instance-level detectability, particularly for thin and

low-contrast cracks. Importantly, removing morphological operations does not substantially deteriorate the quality of extracted contours for vectorisation, but it does preserve significantly better segmentation metrics.

Based on these findings, we revised the methodology and adopted the **V0_raw** configuration (no morphological post-processing) in the final pipeline. This decision prioritises metric fidelity and faithful network behaviour over aesthetic GIS smoothness, resulting in a more honest and deployment-oriented evaluation.

Interpretation of “lightweight” vector representations.

We thank the reviewer for requesting a clearer and quantitatively supported interpretation of the term *lightweight*.

In the revised manuscript, we clarify that *lightweight* refers to the vector nature and GIS compatibility of the representation, rather than to an explicit optimisation of polygon complexity or storage size. To support this interpretation with quantitative evidence, we added an analysis of geometric complexity based on polygon vertex counts in Section 5.2.

Specifically, the manuscript now states (Page 27, Paragraph 4, Lines 793–797):

“Beyond visual inspection, the geometric properties of the extracted vector representations provide insight into the structural complexity of the detected defects. Table 11 summarises the geometric complexity of all predicted defect polygons, measured as the number of vertices per instance. The median polygon consists of 51.5 vertices, indicating that most defects are represented by relatively compact geometries.”

This analysis provides an explicit and interpretable measure of representation complexity, showing that most defects are described by moderately sized vector geometries, while a small fraction of highly elongated cracks yields a long-tailed distribution. No polygon simplification is applied, ensuring that geometric fidelity is preserved.

Accordingly, the term *lightweight* is now used strictly to denote a compact, interpretable, and GIS-ready alternative to dense pixel-level masks, rather than an aggressively optimised or compressed representation.