

Reviewer 2

Comment: “Figure 1 still presents an instance filtering step based on IoU with a threshold of 0.50 as part of the pipeline. This contradicts the textual explanation stating that the GIS output is generated from the complete set of predictions without any ground-truth-dependent selection.”

We thank the reviewer for pointing out this inconsistency between the original pipeline diagram and the textual description.

In the revised manuscript, Figure 1 has been redesigned to clearly separate deployable inference stages from evaluation-only procedures. All ground-truth-dependent operations (including IoU-based instance matching) have been explicitly removed from the main pipeline and are now confined to a clearly labelled evaluation branch.

The updated figure now shows that the GIS output is generated from the complete set of predicted defect instances obtained via connected-component analysis, without any filtering or selection based on ground truth. IoU-based matching is shown only as part of the instance-level evaluation stage and is visually and conceptually decoupled from the deployable pipeline.

Updated elements:

- Figure 1: IoU-based filtering removed from the main pipeline; evaluation-only operations are explicitly separated.
- Figure caption and accompanying text updated to reinforce that no ground-truth-dependent selection is performed during inference or GIS export.

Comment: “Related to this issue, the use of wording such as *true positive mask* in the illustrative pipeline section is problematic. Even if the intention is merely to show an example that happens to correspond to a correct detection, this phrasing implies that true-positive selection is an actual processing step within the pipeline.”

We agree with the reviewer that the original terminology was misleading and could be interpreted as implying a ground-truth-dependent processing step.

In the revised manuscript, all occurrences of terms such as *true positive mask* have been removed. Illustrative examples are now explicitly described as *representative predicted instances* selected post hoc for visualisation purposes only.

The text has been rewritten to clearly state that these examples are chosen solely to illustrate the contour extraction and vectorisation process, and that the pipeline itself operates on the full set of predictions, independently of any ground truth information.

Updated wording example (*Section 5.1, Page 27, Paragraph 1, Lines 811-812*):

“The resulting *predicted instance mask*, shown here for illustration, serves as the basis for polygon generation and subsequent geospatial mapping”

Comment: “The manuscript also requires stronger consistency regarding post-processing. The paper states that the final configuration relies on raw thresholding without post-processing, yet the contour extraction section still applies a minimal area threshold to remove small connected components.”

We thank the reviewer for highlighting this important consistency issue.

In response, we have explicitly clarified the role of minimal area filtering and aligned it with the ablation terminology and the final selected configuration.

First, (*Section 3.9, Page 13, Paragraph 2, Lines 465-471*) now includes an explicit clarification:

”A minimal area threshold $A_{\min}$ is applied to remove degenerate components arising from isolated noise pixels. This operation is not intended as a detection refinement or confidence-based filtering step, but solely as a geometric validity safeguard. Without this filtering, a large number of extremely small connected components fail to produce valid contours or polygonal geometries, which would compromise the robustness of the vectorisation and GIS export stages. No morphological operations or ground-truth-dependent selection are applied at this stage.”

Second, (*Section 5.1, Page 28, Paragraph 4, Lines 818-822*) now includes a dedicated paragraph titled “*Role of minimal area filtering in polygon generation*”, together with Table 12, which quantitatively evaluates the effect of this filtering step:

”Table 12 compares the number and quality of polygonal representations obtained with and without area filtering. Without filtering ($A_{\min} = 0$), a substantial fraction of connected components fails to produce valid contours, leading to missing or degenerate polygon outputs. Applying a minimal area threshold eliminates these failures and results in a fully valid GIS layer without altering the underlying segmentation results.”

These additions ensure full consistency between the methodological description, the ablation terminology (V0_raw), and the deployable pipeline, and make explicit that pixel-level segmentation remains raw, while minimal area filtering is applied only to guarantee geometric validity during vectorisation.

Comment: “The added stability analysis is useful, but the conclusions drawn from it are currently overstated. The selected operating threshold varies substantially across different test set sizes, and the

analysis does not capture variability due to training randomness. If the term *robust* is to be used, the analysis should include multiple random seeds or repeated splits; otherwise, the conclusions should be softened.”

We thank the reviewer for this important observation. This comment directly motivated a substantial extension and restructuring of the stability analysis in the revised manuscript.

In the original revision, stability was assessed only with respect to the size of the test set, using a single trained model and a fixed random seed. As correctly pointed out by the reviewer, this analysis did not capture variability induced by training randomness and therefore did not justify strong claims regarding robustness.

In response, we have conducted a new experimental study in which the complete training, validation-based threshold selection, and test evaluation procedure is repeated using five different random seeds. This additional analysis explicitly quantifies how stochastic factors during training affect both the selected operating threshold τ^* and the resulting segmentation performance.

As a result, Section 5.2 has been restructured into two complementary analyses: (i) an analysis of performance stability with respect to test set size, using a fixed trained model; and (ii) a new analysis of sensitivity to training randomness across multiple seeds.

Importantly, the revised manuscript now clearly distinguishes between variability in the calibrated threshold and variability in the resulting performance metrics. While the validation-selected operating threshold τ^* is observed to vary across both test set sizes and random seeds, the corresponding test-set Dice and IoU values remain confined to a relatively narrow range.

Excerpt from the revised manuscript (Section 4.1.2, Page 17, Paragraph 5, Lines 599–601):

“Although the validation-selected operating threshold τ^* varies across random seeds, the resulting test-set performance remains stable. Dice values are confined to a narrow interval despite noticeable differences in τ^* , indicating that the proposed segmentation pipeline is not critically dependent on fine-grained threshold tuning within the explored operating range.”

The results of this new analysis are summarised in Table 5, which reports the selected threshold and corresponding test-set metrics for each random seed. All conclusions in this section have been carefully softened to reflect the scope of the evaluation, and the term *robust* is now used only in reference to the observed stability of performance metrics, not to the threshold value itself.

We believe that this extended and restructured stability analysis directly addresses the reviewer’s concern and provides a more accurate and transparent characterisation of the variability inherent to both data sampling and training randomness.