Review for "Towards automated identification of mass movements in spaceborne interferograms: Comparing expert mapping and deep learning approaches"

This manuscript discusses mapping mass-movements in alpine environments in SAR interferograms. They focus on two aspects: the first being the variability among human experts mass-movement phase patterns and the second being the automatic mapping of the same interferograms by several deep neural networks, including a modification of the well-known UNet++ architecture (Attention UNet++). For this purpose, four Sentinel-1 interferograms with temporal baselines of 12 days (three) and 18 days (one) of the Valais canton in Switzerland were computed. They find a large variability in the human-mapped mass-movements of one interferogram ranging from 0.18 - 0.49 IoU. After evaluating various convolutional neural networks, UNet++ with the ResNet-18 encoder was chosen based on a very extensive sensitivity analysis. The performance of this network falls within the range of variability of the experts.

This manuscript is pertinent to land-deformation studies, natural-hazard research, and the broader InSAR data user base that intends to use deep learning for downstream segmentation tasks. While I appreciate and commend the authors on their methodological thoroughness, I see two major points that need to be addressed before the work can be published. First, this is a very long manuscript. I recognize the effort put into organizing it into many, many subsections. However, I think the length makes it conducive for a reader to lose the track(s) of the study (as I myself did several times). For this reason, I would recommend keeping the sensitivity analysis out of the main manuscript. Second, I agree with the concerns raised by referee 1 (https://doi.org/10.5194/egusphere-2026-375-RC1) regarding the exclusion of negative samples from the training dataset and the weak justification for this in the discussion. In addition to these, there are a few specific comments and technical errors which also need to be addressed. I have listed them below.

Specific comments

1. Line 182: How do make this assumption when it is impossible to know the impact of training a neural network in a supervised way with "missing labels"? What do you mean to convey here?

2. Line 188: How do you still have training samples with velocities starting from 10 cm/a when they are supposedly filtered out as mentioned in line 186?

3. Line 357: You state that the full dataset consists of 'centre-based' sampled patches while having stated earlier that both negative and positive patches were sampled with a sliding window (Section 2.3.4, lines 215 - 220) and that the full dataset contains both the positive and negative patches (lines 226 - 227). Are the negative data samples constructed with 'centre-based' sampling? In Table 3 there is no mention of another dataset. The sentence 'Minor decreases compared to models trained on centre-based sampled data...' implies the opposite of what the metrics show in Table 3. Please rephrase this sentence to make it clearer.

4. Section 4.2.1 contains important information as well as repeating information already mentioned in Section 2.3.2 about the labels used to train the neural networks. Please consider merging the redundant points into Section 2.3.2.

5. Lines 500 - 504: Is the evaluation on the test set from Bralet et al., 2024 being alluded to in Section 3.2.1 as the 'centre-based' sampled dataset? Why is it mentioned only in the discussion and not in the results? Also, simply stating that 'performance dropped by approximately 11%' when running inference on a test dataset containing negative samples is vague. Which metric are you referring to? In a similar vein stating "When predicting over a complete raster, such as the test area in Queyras Park, the performance degrades even further" is not enough to support your claim. I agree with the comments and suggestions from referee 1 regarding the exclusion of negative samples from your training dataset.

6. Lines 514 - 517: I would also mention the other metrics used for assessing the semantic segmentation in Appendix E.

7. Lines 530 - 533: Could you please provide this information as a proportion/percentage of the total number of samples? Also, what is the slowest mass-movement captured by SAR that is detectable by a human mapper? How many such slow

events actually occur in nature? Answering these questions could help in figuring whether it is even a problem that needs to be addressed.

8. A recent review article published in this journal (https://doi.org/10.5194/nhess-26-487-2026) discusses the use of RNN's, LSTM as well as transformers, particularly to make use of the temporal dimension. You mentioned that it is standard practice for domain experts to use a time series to make sure what they delineate is actually a mass-movement (Lines 458 - 460). I would suggest adding a few sentences about this in either Section 4.2.3 or in Section 5.

Technical comments

1. Line 24: This is a persnickety comment, but repeat acquisitions already implies that they are over the same area at different times.
2. Line 25-26: "multiple/multi-pass (ascending/descending) acquisitions and derives...". And I presume you mean displacement time series/ time series of displacement rates?
3. Line 28: Landslide state of activity.
4. Line 29: It might be worth elaborating the factors/reasons for poor coherence in mountainous areas.
5. Line 34: "...when PSI does not." is an incomplete sentence.
6. Line 47: Add a comma before the citation and remove parenthesis around the citation.
7. Line 55: Change the apostrophe in area and elevation numbers to a comma. Please make this change for other numbers in this paper, where applicable
8. Figure 1: While I can appreciate the authors trying to convey a lot of information in a summarized way through this workflow figure, the subsection numbers above each process is distracting. I would recommend removing these, also because I mentioned in the specific comments that there are too many subsections and sub-subsections fragmenting the paper. Some more comments to this figure:
   (a) Please remake the figure with at least 300 dpi.
   (b) Please rename "Setup" in the Domain-expert variability block (and the corresponding subsection title) to something more descriptive. I would choose another interferogram to show the delineation part, maybe something without shadow?
   (c) Please remove the sub-panel labels '3a' and 'b' from Evaluation of expert's output block.
9. Line 62-72: There is a switch between active and passive voice from sentence to sentence. Please stick to one voice.
10. Line 74: 'In the following subsections, ...'
11. Line 81: Sentinel-1 has 6 day repeat globally when two satellites are active.
12. Figure 3:
    (a) Please use a different color for the polygon indicating the "Setup" areas.
    (b) I and perhaps other readers would be interested in knowing what exactly are the different patterns in these AOIs? In lines 99-101 you make a distinction between sites 3-8 and 8-12. Maybe you could mention the different mass-movements as subtitles of the sub-panels?
13. Figure 4: the cp and icp polygons on the interferograms are indistinguishable.

14. Line 203: Reference the sub-panel related to the folds.

15. Line 344: Fig. 10 (and most other figures) appears much later in the manuscript which means a lot of scrolling back and forth. Please place your figures as close as possible to the subsection where it is first referenced!

16. Table 3: Is the Hausdorff distance measured after geocoding the identified mass-movements polygons? Are the units then in meters?

17. Figure 8:

    (a) Same comment as Figure 4 regarding the visibility of cp and icp delineations.

    (b) You reference Fig. 8 in line 419 and state that the DL segmentation masks show general agreement with the expert mappings. Where are the DL segmentation masks shown?

18. Line 484: "very temporal?"

19. Line 496: "external dataset" (singular).

20. Line 517: ":augmentation" (lowercase following the double colon).