Review of ‘Refined glacial lake extraction in high Asia by Deep Neural Network and Superpixel-based Conditional Random Field’

This manuscript proposed a refined methodology to extract glacial lakes from satellite imagery. Overall it is reasonably clear and well-written. Comments are below

Main comments

- Page 2 lines 39-44 - The discussion about semi-automatic methods is unclear. Please point out what makes these methods semi-automatic. For example, are these methods that use specified thresholds or feature as part of their workflow? Why are they more regionally restrictive than automatic methods. Please revise this part.
- Page 2 line 56 - Why is the method in Zhao et al. 2018 only applicable to Landsat images (is this due to the bands available, if so, please state this as it is more specific).
- Page 2 line 67 - ‘with NDWI as the spatial attention’ - what does it mean to have a feature as the attention mechanism? Or do you mean something different here? I could not find this in the paper (He et al. 2021). Did you mean to reference the paper *’J. Wang, F. Chen, M. Zhang and B. Yu, ”NAU-Net: A New Deep Learning Framework in Glacial Lake Detection,” in IEEE Geoscience and Remote Sensing Letters, vol. 19, pp. 1-5, 2022, Art no. 2000905, doi: 10.1109/LGRS.2022.3165045’*?
- lines 124 -125 I usually use data, like Landsat, downloaded directly from e.g. NOAA or similar sources, and am not very familiar with Google Earth Engine. Other readers of this journal may be similar. I don’t follow specifically what images are used or what the different levels (e.g. levels 14 to 19 of Google Earth images) mean - could you please elaborate?
- line 134 Could these image tiles randomly selected for validation be part of the same scene as those used for training data, and if so, how do you think this may impact your scores?
- lines 147-148 - For this audience I think more information is needed as to what SLIC and Dense CRF are, and why they are chosen.
- In Figure 3 - Are the arrows that indicate ‘First optimization ’ and ‘Second optimization’ steps (e.g., in a code) or is the first optimization the steps from labels/GEE images to output 1, and if so would it be better to put a box around this part and call it ‘first optimization’. Similar comment applies to the ‘second optimization’.
- lines 162-165 The reader might wonder what is so special about using LinkNet. It’s a little hard to follow the motivation here - LinkNet is able to extract deep/complex features with fewer weights than a standard U-Net - is this essential for this problem? Or does it have to do with the addition property of LinkNet when the features from the encoder and decoder are combined?
- Figure 4 - tell the reader what the difference is between a deconv is in comparison to a transposed conv and why you have this distinction in your network.
- Both sections 3.2.1 and 3.2.2 don’t give the reader an idea as to why a superpixel segmentation algorithm is needed, or why a conditional random field are chosen. This information is also not clear earlier (in introduction or background material). Hence it’s hard to read these two sections and extract useful information. If these details aren’t discussed further, the overview (i.e., why these two methods are chosen) could be put earlier in the manuscript, while the details in 3.2.1 and 3.2.2 could be put in an appendix.
- lines 230 ‘full-text’ information - can you place this in the context of the present study?
- No details about RF or SVM are given. We need more information to understand if this is a reasonable comparison. It these two methods were used in a previous study by the authors then they could refer to this here. These two could also be omitted since there is also a comparison to

UNet and EfficientNet UNet if adding that information would make the manuscript too long and the emphasis is not on feature learning vs feature engineering.

- lines 271-273 - Show in your figures where the false detections and shadows are. For example, figure 5 does not say what the red and yellow circles are referring to. A similar comment applies to the multi-color circles in Figure 6.
- I am not sure if contribution (2) on page 3 is considered a contribution since this loss function has been used in other studies. However, it may be the first time this has been used for segmentation of small objects? For example, a related (but different) approach was taken for the problem of sea ice floe segmentation ('Nagi, A.S.; Kumar, D.; Sola, D.; Scott, K.A. RUF: Effective Sea Ice Floe Segmentation Using End-to-End RES-UNET-CRF with Dual Loss. Remote Sens. 2021, 13, 2460. https://doi.org/10.3390/rs13132460') but without the Lovaz loss, which seems to make a significant different here. It might be worth pointing this out (perhaps in the conclusions) since the small lakes are somewhat similar to ice floes in that they are irregular small objects in background state.
- the conclusions refer to the high F1 score in 'the study area' but not (more generally) the results in the QNNR - why is this?

Minor comments

- In the abstract the authors refer to 'Google Earth' images - please be more specific about what images are used. In addition two different image resolutions (2.11 m and 0.52 m) are referred to but it's not sure why two different resolutions are used. They also refer to 'the study region' and then later (in the abstract) 'Qumolongma National Nature reserve', which adds to the confusion.
- abstract line 20 - if pixel spacing is 2.11 m then wouldn't 6 × 6 pixels be smaller than $160m^2$.
- page 2 line 1 'time and vigor' - A different word could be used than vigor - resources?
- page 2 line 62 'EfficintNet' typo
- page 2 line 63 'better result' - better than what?
- page 2 line 69 'area difference between positive and negative samples' - do you mean the difference between the area occupied by positive and negative samples (with there being more area for one than the other)?
- line 98 - remove 'Besides' before 'no large rivers', for example change to 'There are no large rivers in the study area'.
- It would be helpful to show the QNNR area in Figure 1 and then add text to the figure caption indicating which area is used for train/test and how the evaluation for the QNNR is done different that the entire study region (if I follow correctly only inference was done for the QNNR).
- The inset in Figure 7 is far too small.
- line 366 - I am not sure the reader will know what the 'single-variable' method is.