

This paper describes a new convolutional neural network model for extracting waterbodies from Earth observation imagery. It is an interesting method that employs attention to improve on existing work.

I have some concerns about the method I would like addressed. In particular I think it would be hard to replicate the current work from the paper alone, so any improvements to the description of the method and experiment would be welcome. I also have several smaller comments.

Major comments:

- I find the introduction very hard to follow. It goes deeply into discussion of complex and mostly irrelevant literature. I would be much happier to see a discussion of *waterbody extraction* methods rather than an in-depth description of the history of deep image segmentation methods. Improving this would greatly improve the paper.
- The figures would be improved by the addition of layer/weight dimensions, e.g. Figure 2 would be much clearer if the concatenation axis was specified in this way. Same for Figure 4. This would make the paper easier to reproduce.
- Please provide a link to the appropriately licensed code used in this work, so that it can be examined, reused, and reproduced.
- The method section implies that Landsat 8 was supersampled to 15m using the panchromatic band. This is not an obvious method and there are lots of ways to do it. At minimum a citation to the method used would be required; I'd like a description of what was done. How are the other bands "fused" with the panchromatic band?
- There is little to no discussion of labels. How are the labels formatted for the (supervised) training and prediction? Where do the labels of the Google and Landsat 8 datasets come from and are they in the same format? Etc.
- There is ambiguity about the comparison models. A lot of this paper stands on the fact that HA-Net outperforms other models, but there is no description of a) whether these models were retrained on this dataset and b) if so, how were they retrained. This could greatly impact results, and so is very important.
- Tables 1,3-7: Are these differences statistically significant? I normally would not ask but training seems to be very fast, so it would be possible to do a Monte Carlo or bootstrap analysis. It would be nice to see uncertainties on the performance metrics.
- Do you upsample Landsat 8 to match Google? If not, how do you adapt your model to work on Landsat 8? This is not clear from the paper.

Minor comments:

- Abstract
 - o What is the "Google dataset" and the "Landsat-8 dataset"? These are not defined and not commonly understood.
 - o HA-Net is not defined. I assume that this is your model, but you should make it clear on its first use.
 - o What is "TWR"? Not defined and not in common parlance.
 - o "...which proves to be more robust than other advanced model." I don't understand this sentence – what proves to be more robust, what does robust mean, and what is the other advanced model?
 - o HA-Net doesn't achieve "optimality", it's just better than the other models at the metrics. "optimality" is quite specific mathematically. Similarly HA-Net doesn't achieve "optimal overall accuracy". This is an issue throughout the paper.
- Introduction
 - o L29: "0.83%/a" -> "0.83%/year"? I assume "a" stands for "annum" but I'm not sure.
 - o L31-33: I think the connection between "lakes are important" and "automatic extraction of lakes is important" is not clear yet at this stage of the introduction. Consider moving this sentence to after the introduction of remote sensing.

- L39: It is not the need for analysis that increases with increasing data, but our ability to perform that analysis.
- L41: DCNNs do not have greater robustness than traditional methods. Perhaps they have higher accuracy on image tasks – in which case say that. Robustness is not synonymous with accuracy.
- L54: RNN and CNN are undefined.
- Introduction feels disordered. Consider rearranging so it flows better.
- L87: What is the “boundary ambiguity problem”? This is not defined.
- The link between the literature reviewed and the current work is not so obvious. A lot of the literature review is quite detailed – maybe could be moved to related work to keep the introduction on topic.
- Related work
 - L161-165: ResNet is a DCNN, but this paragraph asserts that different authors introduced dilated convolution into ResNet and into DCNNs.
- Method
 - L246: “HS-GCN” is undefined. I assume it means “hybrid-scale global convolution network”
 - L252: “...can effectively suppress the noise in remote sensing images.” This is an unsubstantiated claim. It may well be the motivation for its use, but no evidence is given that it’s effective in this way.
 - Figure 3 caption: What are k, c, s, w, h, BN?
 - L260-261: “GCN-based... tasks.” is not a sentence, something is missing here.
 - Missing citation to PixelShuffle in Section 3.3.
- Experiment
 - L299: It seems odd to say that Landsat 8 “will provide [data] for at least five years”. The satellite has been up since 2013, so it has already been providing data for 8 years.
 - L301: Why only four images, and why these four in particular?
 - Figure 7: “flap” should read “flip”.
 - How large is the augmented training set?
 - What was the ResNet-34 model pretrained on?
 - 4.2.2 is good, nice to see clear description of hyperparameters. Was there any hyperparameter optimisation or were these set arbitrarily?
 - L327: “the weights converged to the global optimum” - how can you confirm that they converged to the global optimum? Maybe you mean local? What was the stopping criterion?
 - L335: I think that the True Water Rate is perhaps more commonly called the precision, and since you are using recall too it may make sense to call TWR the precision.
 - L336: What are the k categories? And if there are k categories then how are positive/negative to be interpreted in these definitions?
 - 4.3.1: Did you retrain the models being compared or were they originally trained on water classification? I’m not sure that this is a fair comparison unless they were trained on the same task. MSLWENet, MWEN, and SR-segnet seem to be water classifiers but maybe you could provide some information about what they were trained on? Was it also the Google dataset?
 - L358, L457: “achieves optimality” - as before
 - Table 2 seems to imply that you did train the other models as well. Please provide details of this.
 - Table 2: “complexity” means “number of parameters” or “number of trainable parameters”? Please define appropriately.
 - For these evaluation metrics labels are required. Where are these labels sourced from? Is this also the Google dataset? What format do they have? Please explain.
 - I assume that the Figure 9 images were chosen by hand to highlight performance – this is fine but please make this explicit. Tables 3 and 4 may be somewhat misleading since these locations are hand-selected, so please make the captions of those tables explicit.
 - What are the yellow ellipses in Figure 9? These aren’t explained anywhere. I think they are meant to guide the eye, but I find them quite distracting. Consider removing.

- Table 5: Were these results obtained on the training and testing set? If so, then this is a form of model selection and invalidates the testing set. If they were instead tested on a subset of the training set, this would be OK.
- 4.3.5: An ablation experiment is excellent here. I like it.
- What is the spatial resolution of the Google dataset and how does it compare to Landsat 8?