The Cryosphere Discuss.,
https://doi.org/10.5194/tc-2020-310-RC3, 2020

Interactive comment on “Image Classification of Marine-Terminating Outlet Glaciers using Deep Learning Methods” *by* Melanie Marochov et al.

Anonymous Referee #3

Received and published: 22 December 2020

Marochov et al. develop a two-stage machine-learning pipeline to automatically segment glacier calving fronts into seven distinct classes. The initial phase of the pipeline uses a VGG16 convolutional architecture to label whole tiles as one of the seven classes (using a fully-connected layer at the end). Phase two uses the output of the initial labeling to perform pixel-level classification of the landscape into the seven classes. The authors explore a range of training regimes and find state-of-the-art performance for multi-class segmentation of glacier calving fronts. I believe this manuscript provides timely and suitable results for the community. However, there remain a number of major and minor issues that need to be addressed before the manuscript can be considered for publication:

Major Comments:

- There needs to be more justification as to why a two-stage pipeline is necessary. What happens if you directly start with a pixel-level classification of the features? It seems to me the point is that by using a pre-trained VGG16 to first classify the tiles and then using those classification as the training for the second phase, you are cutting down on the amount of required training labels to directly train on pixel-level classification. Is that true? And is this really the only reason? This needs to be communicated better.
- Statements regarding generalizability:
  - Lines 118-119: “they are also applicable to mapping outlet glaciers anywhere in the world, including Antarctica”
  - Lines 998-990: “once the phase one models are trained and weights are saved, no further training is required to apply the workflow to other marine-terminating outlet glaciers.”
  - Lines 1018-1019: “our adapted CSC workflow is transferable and capable of maintaining a high level of performance on other unseen outlet glaciers in Greenland and likely other glaciated regions such as Antarctica.”

  These statements need justification. You haven’t shown how this pipeline performs in outside of Greenland, and delineating calving fronts in Antarctica can indeed be very different due to differences in the glacier sizes and mélange. In fact, Helheim glacier is a rather non-representative area given the shape of the calving front and the fjord. I recommend testing the pipeline in more out-of-sample areas including major fjords like Jakobshavn that are both important and have significant independent studies over time for context.
- Table 1: As mentioned in the caption, there is some class imbalance due to the geographical extent of the different features. You mention you tried to even out the imbalance in the selection as much as possible, but how is the remaining

imbalance dealt with and how is this affecting your results? You mention this in section 5.4, but the study will be significantly improved if the class imbalance is addressed. Can this be addressed in the construction of the loss function? Do you think it's not necessary or that the results won't be significantly affected to justify implementing a more nuanced loss function?

- Lines 348, 379-380: You mention the training hyper-parameters were kept constant for all nine model variations. But no justification is provided that the different models would require the same hyperparameters for optimal training. In fact, one would expect the different models to have different requirements. If this is the case, then comparing the model performances is not fair given that they have not all been optimally trained.
- Figure 4: In phase-two, instead of doing pixel-level classification for one patch at a time using an architecture with a fully-connected layer at the end, why not use a fully convolutional architecture that provides a different class for each pixel directly and convolves over the whole image, skipping the need for feeding in patches?
- Section 4.1: when reporting the performance of the pipeline with respect to manually classification, it is important to also report the uncertainty associated with manual classification at the pixel level. The exact boundaries of surfaces may differ from person to person, but this is not discussed in your performance evaluation.

Minor Comments:

Lines 89-92: The comparison between previous efforts and the conclusion that the Baumhoer study has the most accurate results seems unjustified. These studies are each on different geopgraphical areas and use different input data that have different resolutions. The delineation accuracy of the models in distance units (meters) is not

a fair comparison when the inputs have different resolutions. Instead, a pixel-based comparison seems more appropriate here. Also, it is a bit confusing why the statistics of the test data are reported for the first two studies, but both the test and training statistics are reported for the Baumhoer study (this confuses the reader in terms of which numbers should be compared).

Line 114: training size of 13 images. But how large are these tiles in terms of pixels? The number by itself doesn't convey any information.

Line 116: "resulting class predictions are then used as training data specific to the unseen input image". This sentence is hard to understand. I'm assuming you mean that the classification of the first stage is used as training in the second stage, even if regions where the first stage wasn't trained on. But it needs to be stated more clearly.

Line 311: "The image tiles are then reassembled to creates a class raster which is used as training data for the second model in phase two". Again this sentence is hard to understand until further on into the paper. If you explain what you mean by class raster earlier, it will be much easier to understand.

Line 325: "[. . .] typical of CNN architectures" This is not true. Not all CNN architectures have a fully connected layer. In fact that's what sets fully-convolutional networks like U-Net apart.

Line 341: "only the weights in the final layers of the NN are retrained". Which layers exactly are retrained? How far back does one have to go in the layers to adjust the pre-trained network?

Line 414-415: how does removing tiles with mixed classes (with a 95

Line 426: "new satellite image of the training area". Are you stating that the validation data used during training is never used as one of the actual training images, and therefore it is a more stringent test? But are you using the same "unseen" validation image in every epoch? This needs to be explained more clearly. If the validation loss is being tracked for stopping the training, then it's in essence a part of your training and is different from a novel image during the testing stage that was never used in training as either training or validation data.

Line 428: Do the reported numbers of tiles include additions due to augmentation? Needs to be stated more clearly. If this is the case, then it may not be a fair comparison to the aforementioned studies in terms of comparing the volume of input data as they may not report the numbers after augmentation.

Lines 877-878: Comparing the number of training images to the previous studies here is not a fair comparison because the image sizes are different. In addition to the number of images, it is important to also mention how big (in terms of pixels and resolution) these images are. How do your 13 images compare to e.g. the 38 images of Baumhoer et al. or the 123 images of Mohajerani et al.?

Line 923: "Furthermore, the U-Net architecture will learn shapes that have a limited variability of both form and scale". This statement is not justified, and arguable false. Better justification and citation is needed for such a strong claim, especially given that U-Net has been successfully used in many contexts across many scientific fields from biomedical imaging to glaciology.

Line 955-957: "It is interesting to note that the transfer learning technique benefited from using a larger number of smaller tiles compared to the preferred smaller number of large tiles for the fully trained CNN." Is this because pre-trained networks are less versatile in learning more diverse and spatially connected features across a larger spatial domain? Maybe you can explore and explain this relationship better.

Technical Comments:

Liens 178-187: This is just a soft suggestion and feel free to ignore, but given the relatively long length of the manuscript and the scope of The Cryosphere, the context behind the biological inspiration of neural networks may be unnecessary here.

Line 192: "series of layers containing solutional, non-linearity, and pooling functions". Technically non-linearities are included as part of the convolutional layer given the activation function, whereas the pooling layer is a separate layer.

Line 199: In addition to differences in orientation, you can also mention pooling helps with translational invariance

Line 1052: Missing parenthesis at the end of list of references.

Interactive comment on The Cryosphere Discuss., https://doi.org/10.5194/tc-2020-310, 2020.