Dear Authors

I have received feedback from two reviewers who have reviewed the first and second rounds of your revised manuscript. One reviewer is satisfied with your latest revision. Another reviewer is still sceptical about the novelty of this work. Both reviewers point out some comments that would require the authors to make minor revisions to the manuscript. I accordingly invite you to make further improvements to your manuscript, please pay particular attention to the novelty of your work. Then, I would be willing to accept your manuscript for publication.

Best regards,

Bin Cheng

We are grateful to the editor for the opportunity to improve our manuscript. We have addressed reviewers' comments point by point and revised our manuscript. Please find our responses below.

==================================================================

Reviewer 1:

After the last round of peer review, the author made significant revisions to the paper, resulting in a noticeable improvement in overall quality. I am generally satisfied with the author's response and revisions. However, I still believe that the paper lacks innovation. While the author argues that their method has some relevance for future research in ice floe segmentation, I find the method itself lacking clear scientific significance and methodological value. The author's substantial work primarily involves labeling samples using existing methods and then training a deep learning model for classification. The improvements to the method itself are limited, and it can be said that the main contribution lies in the application of deep learning methods to ice floe segmentation, although this application does hold a certain level of importance in the field.

We appreciate the reviewer's careful review and valuable comments on our manuscript. With regards to the novelty and innovation of the study, this lies in the application of our floe mapping approach to widely available, and free to access, near-real time (NRT) high resolution optical (HRO) images from Copernicus Sentinel-2. Previous studies, e.g. have been limited to airborne or declassified military satellite images (e.g. MEDEA used in Denton and Timmermans, 2022, and Wang and others, 2023) or X-band SAR (Ren and others, 2015, Hwang and others, 2017). Neither of these can provide the regional, NRT coverage to make them viable in operational monitoring for maritime safety. We appreciate that HRO is still cloud and nighttime darkness limited, but the benefit of using Sentinel-2 is in combination with other approaches, especially SAR-based classifications, when cloud-free periods allow. We look forward to the darkness limitation being solved with the future Copernicus LSTM mission.

Please find our responses to the specific comments below.

I suggest the following improvements to the paper:

1. Currently, the method comparison is placed in the appendix. I recommend moving it to the main body of the paper to make it more accessible for readers to read and understand.

Thank for reviewer's suggestion. We have moved the entire appendix to the main body of the manuscript.

2. The architecture of the DL has been described. Not having deep knowledge when it comes to the design of deep learning networks, I am wondering why exactly this architecture was chosen? Have the authors performed tests on different ones, or is there some logical reasoning behind the number of convolutional and pooling layers, or does the exact architecture not really matter as long as you have "enough" layers?

Yes, we have performed several tests for each DL model on different numbers of convolutional layers, batch normalization, drop out, kernel size, etc. The performances of the same model under different parameter settings did not improve significantly, while the performances of different DL models varied. Since the currently available training datasets are still limited, it is premature to emphasize the exact architecture which may lead readers to mistakenly think that the current model is robust enough for any complex ice conditions. As training datasets become richer, more robust floe segmentation models will be developed. Specifying an exact model architecture then makes more sense.

3. How does the method perform when dealing with unique sea ice conditions, such as the presence of melt ponds? This aspect should be addressed in the paper.

We are continuing to process sea ice images with challenging ice conditions. But an idea following from this is to potentially use additional sea ice classification methods to detect melt pond regions. Then catalog these melt pond pixels to the class of floe in DL prediction, and finally refine the floe segmentation using the proposed post-processing.

Thank you for the comment. We have added this in the revised manuscript.

================================================================================

Reviewer 2:

The authors have made some solid revisions according to the comments from several reviewers. I think the manuscript has greatly improved and I only have some minor comments this time.

We thank the reviewer for their careful review and valuable comments on our manuscript. Please see our response to your comment below.

1) (Line 130) It is, obviously, convenient to use GVF snake-based method as an “annotation tool” to help automatically label individual ice floes. However, whether automatically labeling would affects the result?

Automatic data labelling would not affect the results since the GVF snake-based method produces good individual ice floe segmentation results from local-scale MIZ images, allowing DL models to be trained on well-labelled dataset.

2) (Line 138) The boundary line consists of two pixels: inner and outer boundaries of floes. Does it mean that the inner boundary consist of one ice pixel and the outer boundary consist of another ice pixel?

The inner boundary consists of pixels from the floe itself. While the outer boundary mainly consists of pixels from ambiguous edges between the floe and another floe or the water.

3) (Line 150) Why “1×1, 1×2, 2×3, 3×4, 4×5 sub-images”, not “1×1, 2×2, 3×3, 4×4, 5×5 sub-images”? Or in other words, what’s the advantages of the former cropping approach?

An overall image was divided randomly into sub-images. Fig. 4 is just an example to illustrate this multi-scale division process. We apologize for the confusion, and we have modified the description.

4) (Line 238) Did the validation (52 pairs) and test (37 pairs) are also be trained? As illustrated in the abstract, is it enough to only take 290 pairs into training for successful results?

No. The validation dataset was used to evaluate how well the model fits the training dataset during training, e.g., to check whether the model is underfitting or overfitting the training dataset. It does not directly participate in the update of model weights. While the test dataset was used to evaluate the performance of final trained model, e.g., to compare the performance between different models.

The DL model in this work was trained on 290 pairs training dataset and produces acceptable results. Thanks for the comment. We have modified the wording in the abstract.

5） (Line 356) How many images the average segmentation time was calculated from, if it could be told? And I think it is also necessary to explain the hardware used. In addition, should training time also be taken into account?

37 test images were used to compare the segmentation time between models. The hardware was Intel(R) Core(TM) i7-4600U CPU $@$ 2.10GHz, 16 GB RAM, Integrated Graphics Card, which was described in Section 2.2 Software and hardware.

The required training time from fast to slow is FCN family, SegNet, U-Net & ResUNet, U-Net++ & ResUNet++. The differences between models are only a few minutes and could be negligible.

6） (Line 364) It was obvious that there were many over-segmentation on ice floes without post-processing as shown in Fig.A3. Does that mean for now there aren't auto approaches to overcome the over-segmentation good enough? In this way, whether post-processing has become a necessary step in the whole processing progress? What's more, will post-processing affect the automation of whole processing progress?

So far, we have not found a good solution to the over-segmentation problem for a single optical floe image, especially for over-segmentations that are ambiguous (e.g., the middle-left part of the biggest floe in S2-2 image in Fig. A3/Fig. 11). A series of images of the same area over different periods or other types of the remote sensing data may be needed to help determine whether the floe is over-segmented.

Post-processing is used to refine the floe segmentation, such as removing "holes" in large floes, preserving floe shape, and separating potentially under-segmented floes that may affect FSDs. This is a necessary step to maintain the integrity of the ice floe, especially when the surface of the floe is noisy, i.e. contains small melt ponds, partially covered by small clouds, etc.

In our post-processing, we have proposed criteria to automatically check whether the detected floes are well segmented to decide whether to perform subsequent morphological opening and closing. The area cut-off threshold *Ta* for finding the potential under-segmented floes and the area threshold *Ta'* for finding floes require a smooth shape are two main parameters in our post-processing that need to be adjusted according to image scale and/or practical application needs, while the solidity cut-off threshold Ts kept constant at 0.85. The pure morphological operations adopted in our post-processing often require extensive manual parameter tuning to segment floes even from a single image, since they operate on the binarized image and relies on how well the edges are detected between connected floes. As the DL method detects more accurate floe boundaries, morphological operations in our post-processing become less dependent on parameters in refining floe segmentation (a disk-shaped structuring element with a radius of 4 pixels was used in the morphological operations), making the entire process more automated.