In this study the authors used a classical image processing technique to label the ice floe samples, and then used these samples for training a deep learning model, which was used for ice floe segmentation. The authors evaluated the algorithm using two types of remote sensing data and compared its accuracy and runtime with other methods. They claimed that this approach can achieve faster processing speed and higher accuracy. Deep learning models have been widely used in remote sensing image processing, but the prerequisite for obtaining ideal accuracy is usually a sufficient amount of training samples. Sample labeling is usually done manually, which often requires a lot oef manpower and time. Using an automatic labeling method to obtain samples has certain advantages.

Although the deep learning method achieved the best results, which was also expected, using an automatic method for labeling a large number of samples and then for training deep learning models is a commonly used approach. The paper did not provide sufficient innovation, whether in terms of methodology or scientific application. It is recommended that the authors focus more on the methodology itself to address the specific technical issues encountered in the ice floe segmentation, rather than simply using samples to train the deep learning models to obtain so-called high accuracy.

We thank for reviewer for careful review and helpful feedback on our manuscript. Please find our responses to your comments below.

General comments:

Using simple methods for automatic labeling of samples and applying them to the training of deep learning models is a common practice, and this paper does not provide enough innovation in this regard. Therefore, I believe that the originality of the paper is relatively limited.

Current methods for extracting individual ice floes and determining floe size distributions from images remain stay on classical approach, which usually require a lot of human intervention and distance away from practical needs.

Although DL techniques have been rapidly developed and successfully applied in a wide range of fields, its application in ice floe segmentation is rarely studied. A main reason limiting the application of DL techniques to ice floe segmentation is the difficulty in obtaining labelled data, which is a very challenging or even impossible task even by domain experts in a manual way. Therefore, this manuscript introduces an approach to automatically label ice floe images, and explores the feasibility of using a small number of labeled datasets to training a DL model for ice floe segmentation and whether the model can be generalized to wider variety of ice floe images. We believe our work is meaningful for the further development of DL in ice floe segmentation, as well as for sea ice studies.

As the authors point out, one of the advantages of this method is that it can reduce the running time. Classical methods for sample labeling take a considerable amount of time. As the number of training samples increases with the further application of the model, the

training time of the model will also increase. If we only compare it with classical methods, this method additionally needs the time for model training. Of course, if we only compare the running time, the deep learning model takes less. But what is the practical significance of shortening the time? Can it be used for some near-real-time applications?

Following the steps introduced in the manuscript to use the classic method for labelling data can reduce unnecessary trial and error time as well as human intervention.

The time saved in processing the data can compensate for the time spent on labelling and training as the amount of data to process increases.

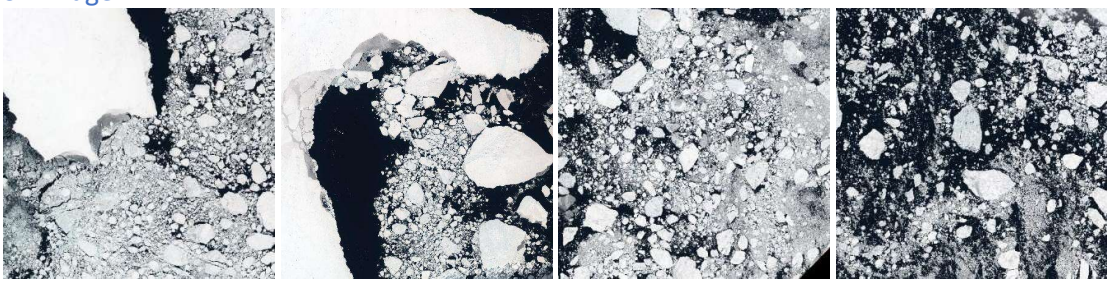
The DL-based method is expected to be applied in marine operations in cold regions such as the Arctic and Antarctic, which require online monitoring of ice conditions in real time and rapid extraction of ice properties to improve maritime safety and provide better data for path planning.

What is the difference between results from classical methods and deep learning methods? The training samples of deep learning come from the classification results of classical methods. If there are some errors in the training samples, these errors may also be introduced into the deep learning model. Although the authors believe that deep learning can overcome this problem by itself, the influence will still exist. How do the classical methods and deep learning methods affect the subsequent acquisition of ice floe parameters, and is the difference obvious?

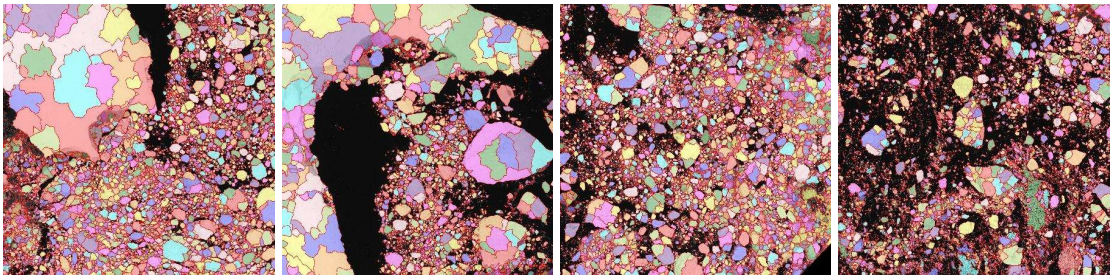
In shortly, the classical method, i.e., GVF snake-based, identifies floe boundaries one by one and takes longer to process images as the number of ice floes increases. The GVF snake-based method also does not well in global-scale floe image segmentation and tends to over-segment big floes.

The DL-based method identifies floe boundaries simultaneously and takes shorter processing time (please see Tab. 5 in the manuscript). Although the DL model was trained on the data annotated by the classical method, it surfers less from the segmentation issue (please see Fig. 12 and 13 in the manuscript, and the figures blow. Please also pay attention to the different colours in the figures which indicate whether the floes are under-/over-segmented). The reason for this, i.e., the explanation of the rest questions in this comment, can be found in our response to your comment ‘‘Line 22’’.

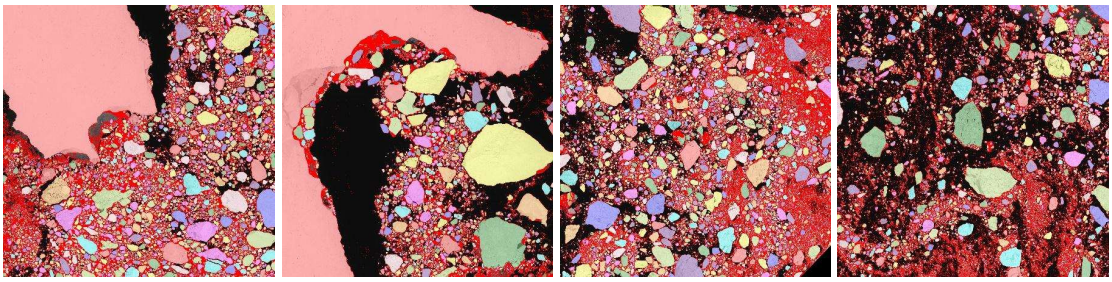
S2 image



GVF snake-based



DL-based



The authors used two resolutions of remote sensing images to test the method, but I did not see the comparison of the two results. Will the spatial resolution have an impact on the algorithm? How does the sample size of different resolutions compare? What kind of impact will it have on the training of the model? How sensitive is the proposed method to the size of the ice floe? Can similar accuracy be achieved in other regions of the Arctic or at other times?

The local-scale airborne MIZ images were used to train and test the DL models, while the global-scale satellite images were only used as additional test data to investigate the generalization ability of the trained DL models from local-scale MIZ images to global-scale images and they were never encountered by the DL models during the training (please also see our responses to your comments ‘‘Line 98’’ and ‘‘I did not see the impact of sample size on the method…’’ for details).

The GVF snake-based method tends to over-segment big floes in global-scale satellite floe image, while the DL-based method doesn’t suffer this ‘‘floe size’’ issue although the model was trained on the data annotated by the GVF snake-based method (please see our response to your previous comment ‘‘What is the difference between results from…’’).

Our approach has been applied to extract the ice floes spatially and temporally from Sentinel-2 images. The method can achieve good results when floe surface is relatively flat with few melt ponds. The melt ponds are often recognized as floe boundary pixels by the method. So floes in summer with many melt ponds are often over-segmented.

The parameter settings of the deep learning model are not clear enough, and the influences of multiple parameters on the results need to be compared to obtain the best training and classification results. The method process is not clear, making it difficult for readers to follow

and implement. The testing images is also limited, which makes it difficult to demonstrate the robustness of the method.

We apologize for the lack of detailed descriptions about the models, and we will add the descriptions and diagrams of the models in the revised manuscript.

The title of this manuscript is ‘‘Towards a manual-free labelling approach for deep learning-based ice floe instance segmentation in airborne and high-resolution optical satellite images’’, and the manuscript mainly aims to introduce an approach to automatically label ice floe images, explore the feasibility of using a small number of labeled datasets to training a DL model for ice floe segmentation and whether the model can be generalized to wider variety of ice floe images. The U-Net++ model used in the manuscript is a suggestion after we investigated different DL models, including tuning parameters (e.g., number of convolutional layers, batch normalization, drop out, kernel size), modifying loss function (e.g., dice loss), and also training and testing other architectures such as GAN (with different discriminators and generators), Yolact, Mask R-CNN in addition to the models mentioned in the manuscript. The investigation of DL models is not the key in this manuscript, but another topic/task after using our approach to automatically generate more training datasets of greater variety. Thus, we don’t think it is necessary to present many detailed model comparison results, but present some comparisons between typical DL architectures in the manuscript.

Regarding test images, it is very challenge to obtain ground truth for comparing DL model evaluation metrics (please see our response to your comment ‘‘Line 22’’) and not practical to present so many images in the paper. It is common to use some typical images to demonstrate the robustness of the methods (please see our response to reviewer 3’s comment ‘‘The amount of data is very restricted…’’).

In addition, this method does have some practical value, and the author can consider making the method model public.

Thanks for considering our work has some practical value.

The matlab scripts for the GVF snake method is publicly available and can be found at: https://www.ntnu.edu/imt/books/sea_ice_image_processing_with_matlab

We will consider making the rest of the scripts, including the python version of the GVF snake method, publicly available after the manuscript is published.

Specific comments:

Line 15: What is the difference between "sea ice" and "floe" here?

Here, ‘‘individual pieces of sea ice’’ refers to ‘‘floes’’.

Sea ice is defined as any form of ice that forms as a result of sea water freezing. It has many types, such as level ice, ice floe, brash ice, etc. An ice floe is defined as a piece of flat sea ice.

Line 15-20: This sentence is too long. It is recommended to rewrite it.

We will make this sentence shorter.

Line 20: What does "environment information" refer to?

Environment information includes sea ice types (brash, slush, floe, nilas, level ice etc.), ice concentration, floe position, size and shape, etc., which can be revealed from ice image for climate studies, safe marine operations, or other purposes.

Line 21: What does "floe parameters" refer to?

Floe parameters include floe area, perimeter, mean caliper diameter (MCD), shape property, etc., which are used to characterize floe size distribution (FSD) or for other purposes.

Line 22: Classical methods have some difficulties in distinguishing connected floes, and these errors may also exist in the training samples of deep learning models. Why can deep learning models overcome these problems?

First, there were not so many errors in the training dataset. The method we used (i.e., GVF snake method) has a good ability to segment individual ice floes in local-scale MIZ images. Also, due to the characteristics of floes in MIZ, the aforehand quality control can help remove images that may have potential severely wrong segmentations in the dataset (please see our response to your comment ‘‘Line 170’’).

Second, a DL model learns the most significant features from a training dataset. Small errors in the training dataset can be determined as outliers during model training. Learning from noisy, restricted, or inaccurate labelled data (i.e., weakly-supervised learning) is desirable and a branch of machine learning techniques that aims to liberate the strong need for expensive data labeling processes.

It should be pointed out that, ‘‘due to the dynamic nature of the scene and the ambiguous boundaries between the water and ice during melting, it is a very challenging or even impossible task to obtain an error-free GT, even by domain experts in a manual way’’ (excerpted from (Chai et al., 2020) in ref.). Therefore, our manual-free labelling approach is more practical than manual labeling for DL-based ice floe segmentation.

Line 50: The poor performance of these models may be due to model structures, other types of models (e.g., the model you used) may resolve this problem.

The instance segmentation approaches rely heavily on the object detectors, they tend to miss floes and always have problem in detecting small floes (objects). So the instance segmentation approaches are not suitable for identifying individual ice floes especially when a large number of them are tightly connected to each other.

This is why we use semantic segmentation approaches to solve floe instance segmentation problem.

Figure 2: The title of the figure is too simple and lacks necessary descriptions.

The product IDs of the four S2 in Fig.2 were listed in Tab. 2 which can tell many information about S2 images, e.g., the product ID of S2-1 image ''S2A_MSIL1C_20210527T144921_N0300_R082_T28XEN_20210527T165730'' identifies a Level-1C product acquired by Sentinel-2A on the 27th of May, 2021 at 14:49:21 PM that was acquired over tile 28XEN during relative orbit 082, and processed with Payload Data Ground Segment (PDGS) processing baseline 03.00. More information about each S2 image (e.g., cloud coverage) can be found by searching their IDs in Copernicus Open Access Hub.

We apologize for the lack of the description, and we will add more descriptions about S2 data and the naming scheme in the text, Tab. 2 and in the caption of Fig.2 to make it clearer.

Section 3: Although you have written a lot in this section, it is still difficult to understand the processes. It is recommended that this part should be rewritten with a clearer logic, so that readers can follow and implement the method. A method flowchart is recommended here.

The overall flowchart of training procedure was given in Fig. 7, and the overall flowchart of inference procedure was given in Fig. 8.

Fig. 4 and Fig. 5 showed the flowcharts s of multi-scale division (which is a part of training procedure) and post-processing (which is a part of inference procedure), respectively.

We apologize for the confusion. We will adjust the structure of the manuscript and revise the Method section.

Line 86-87: Some data or references are needed to support this.

(Toyota et al., 2006) in ref. will be added here to support this as they have mentioned that the threshold for separating connected floes needs to be higher than that for separating ice-water regions.

The reference below will also be added as they used different thresholds, from low to high, combined with morphological operations to segment floes.

Denton, A. A., & Timmermans, M. L. (2022). Characterizing the sea-ice floe size distribution in the Canada Basin from high-resolution optical satellite imagery. *The Cryosphere*, *16*(5), 1563-1578.

Line 98: What is the ratio of airborne data to satellite data in the MIZ image, and will it affect the performance of deep learning model?

The ratio of airborne MIZ data (local-scale) to satellite data (global-scale) in training dataset is 1:0.

The training dataset contains only (local-scale) airborne MIZ data, and the DL models were trained only on local-scale airborne MIZ data.

The global-scale satellite images were additional test data only used to investigate the generalization ability of the DL model from local-scale MIZ images to global-scale images. They were never encountered by the DL model during the training.

We used ''MIZ images'' to refer only to local-scale airborne MIZ images and used ''S2'' to refer only to global-scale satellite images in the manuscript. We apologize for not describing them clearly. We will clarify them in the revised manuscript.

Line 111-112: Is this correct? Does it contain real ice floe boundaries?

An object has two types of boundaries: inner and outer boundaries. The inner boundary is the outline pixels inside of the object, and the outer boundary is the outline pixels in the background surrounding the object.

Both inner and outer boundaries of ice floes were labelled as the third class of floe boundaries in our work in order to balance the classes in the training set and widen the gaps between connected floes.

Line 119: Do you mean that this method may not achieve good results in the local regions?

Line 119 talked about the floes in MIZ are usually of similar size and shape, while the floes in regions other than MIZ are generally of varying sizes and shapes.

Regarding the classical method (i.e., the GVF snake method) we used to automatically label individual ice floes:

1. the GVF snake method has good performance in automatically segmenting floes in local/small scale floe images, especially for MIZ images where floes are of similar size and shape, even though a large number of floes are tightly connected to each other.

2. the GVF snake method may have poor performance in global/large scale floe image segmentation. It cannot balance the segmentation of floes with large differences in size and shape, and tends to over-segment bigger floes (e.g., Fig. 12, 13 and figures we presented in your previous comment ''What is the difference between results from…'') .

3. only local-scale (airborne) MIZ images were used as training dataset to train DL models. Since the size and shape of the floes do not change too much in MIZ images, we have an additional multi-scale division step to create more floes of varying sizes and shapes to increase the diversity of the sample.

Line 129: "an ice floe can be resized into several smaller ones of different scales", do you mean that you create more small ice floe objects by resizing the large one?

Yes.

Section 3.1.2: I don't understand why you divided the images into multiple scales and how to implement it specifically. More details are needed here.

The training dataset contains only (local-scale) airborne MIZ images, where the ice floes vary little in size and shape. So we use a multi-scale division step to increase the diversity of floe sizes and shapes in the training set. This is a kind of data augmentation step.

In the method section, I did not see the impact of sample size on the method. You used two different resolution data sources. What is the impact of the number ratio between them on the method? If this method is applied to other regions, it may be more realistic to increase the amount of satellite data. What kind of impact will it have on the classification results mainly based on satellite images?

We used only local-scale airborne MIZ images to train the DL models, while the global-scale satellite images were used as extra data to test the generalization ability of the trained DL model (please see our response to your previous comment ''Line 98''). It is thus impossible to tell the impact of the number ratio between local and global scale data on the method from this manuscript.

But we agree that it would more realistic if including the satellite data in the training set, and the trained DL model would be more robust. In our next step after work on this manuscript, the global-scale satellite data, where the floes are labelled by the DL-based method introduced in this manuscript, will be added to the training set to obtain a more robust DL

model, as we mentioned in the last sentence in the Conclusion section: ‘‘it can also be utilised as a “higher version” of “annotation tool” and produce more “ground truth” from a wide variety of ice image data sources to further train more robust DL models for obtaining more accurate ice parameters from images.’’

Line 170: So you applied an aforehand quality control on the samples, is this manual-free?

Yes, a pair with severely wrong annotations can be automatically filtered out by using the criteria introduced in the post-processing section for finding under-segmented floes. That is, if the ratio of the total area of the labelled floes that do not satisfy the criteria to the total area of all the labelled floes is larger than a threshold, the annotated pair will be removed.

This aforehand quality control is a trial-and-error process. The incorrect segmentation made by the GVF snake-based method occurred in non-MIZ regions and regions blurred by water vapor. This can be avoided if the training images are all unblurred MIZ images.

We apologize for not stating this in the manuscript. We might put a footnote on line 170 stating this aforehand quality control.

Line 172: Is the sample size too small for the deep learning model?

Learning with small dataset is also desired for DL learning techniques as it can be very challenging to obtain a large number of ground truth, especially for tasks like ice floe segmentation.

The main purpose of this manuscript is not to use large dataset to train a very powerful DL model, which is currently unrealistic for ice floe segmentation currently. But we aim to provide an approach to train a DL model for ice floe segmentation without manual labeling of data, and investigate the generalization ability of the DL model. Our results demonstrate that the model trained on restricted datasets that can also generalize to wider datasets. And the trained model can be further used to automatically label more data to gradually obtain more robust DL models.

Line 217: There are no obvious differences between deep learning models of the same category, and their performances are also similar.

We apologize for the lack of descriptions about the models, and we will add the descriptions and diagrams of the models in the revised manuscript.

Regarding the performance comparison by using:

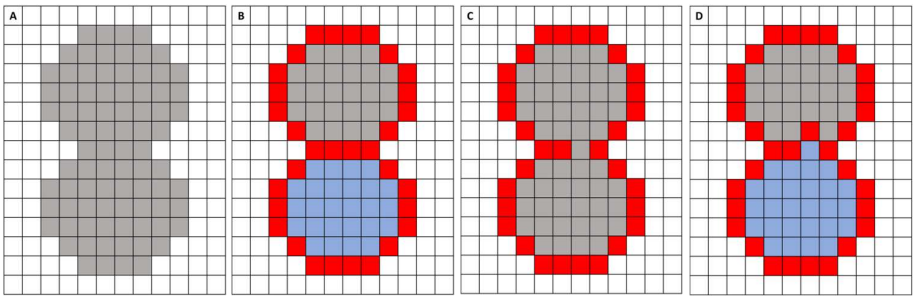
1. Well-known evaluation metrics: they are commonly used for comparing the performances between different DL models. But they are not the only criteria used to evaluate model performance in our work. In addition to requiring extensive manual manipulation to create ground truth, these metrics (also other existing evaluation metrics) have limitation in evaluating the performance of models for floe segmentation (please see a simple example below our response to this comment). Therefore, we used the evaluation metrics as tools to first filter out poorly performing models (e.g., FCN family and SegNet).

2. Visual comparison: this is the most intuitive way to evaluate segmentation results (Chai et al., 2020). We use some typical floe images from low to high ice concentrations to demonstrate and compare the segmentation performance of different methods. The segmented individual ice floes were marked with different colours in Figures 9-10 to indicate which floes were under-/over-segmented. Please pay attention to those colours. We will also add some notations to these figures as suggested by reviewer 2.

====

A simple example to explain the limitation of commonly used evaluation metrics in floe segmentation problem:

A - an image of two connected objects, B - the ground truth (red indicates object boundary), C and D - two segmentation results.



C has only one mis-segmented pixel while D has two mis-segmented pixels, and the performance indicators of C is higher than those of D. But the two objects are still connected to each other in C, while they are detached in D (with 4-connectivity).

Figure 9: It is difficult to distinguish the difference in results between different methods, and the image below is the same.

The segmented individual ice floes in these figures were marked with different colours to show which floes were under-/over-segmented.

Following reviewer 2's suggestion, we will add some notations to these figures, as we did for Fig. 11.

Line 243-244: This may affect the classification performance of deep learning.

Please see our response to your comment ‘‘Line 22’’.

Line 250: The training samples also contain these errors. Why can deep learning automatically overcome this problem?

Please see our response to your comment ‘‘Line 22’’.

Figure 15: Similarly, the title of the figure is too simple and lacks necessary information.

We apologize for the lack of the description. We will write more description about FSD in the figure caption.