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.

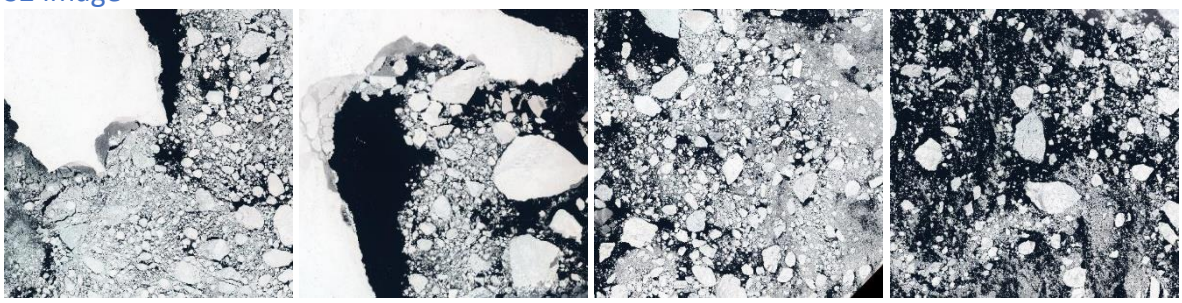
The DL-based method is expected to be applied in marine operations in cold regions such as the Arctic and Antarctic, which require online monitoring of ice conditions in real time and rapid extraction of ice properties to improve maritime safety and provide better data for path planning.

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?

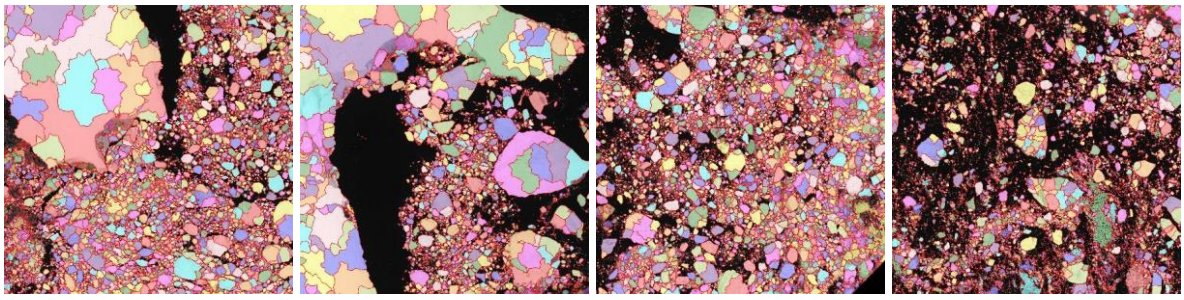
In shortly, the classical method, i.e., GVF snake-based, identifies floe boundaries one by one and takes longer to process images as the number of ice floes increases. The GVF snake-based method also does not well in global-scale floe image segmentation and tends to over-segment big floes.

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’’.

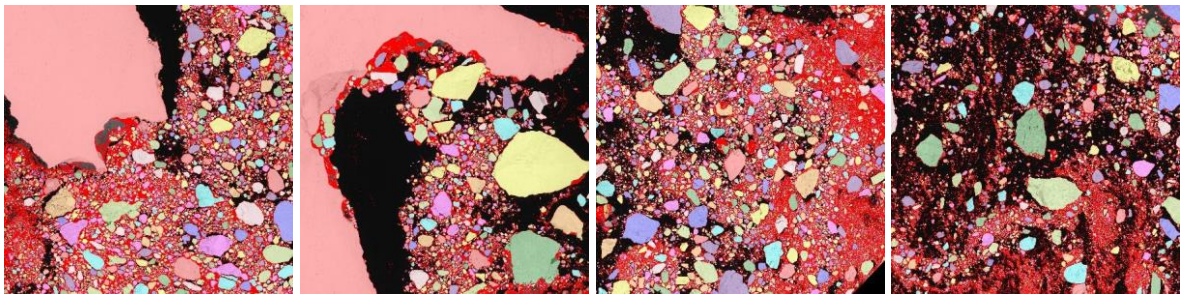
S2 image



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 have modified this sentence.

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 have added 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 have adjusted the structure of the manuscript and modified 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 have added statement about 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.

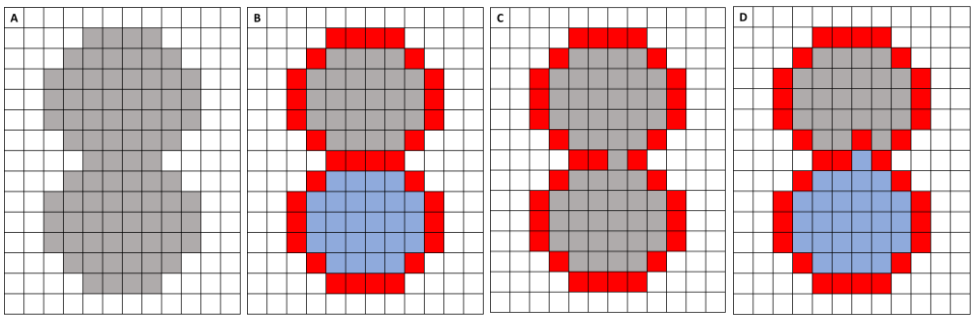
Regarding the performance comparison by using:

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 have added some notations to these figures.

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. Figure 15 has been combined with Figure 16.

Floe size distribution (FSD) becomes a very important parameter in nowadays sea ice modelling; however, high-resolution imagery seems to be the only source to obtain such kind of information. Thus, an automatic image-processing method is also important in this field. This study provided a deep learning-based segmentation method to process airborne and optical satellite images, and obtained good results of FSD. It seems that a completely automatic method to get FSD becomes possible.

Actually, it is not the first time for me to review this manuscript. I understand the solid revision that the authors have conducted to improve the paper. I still encourage the authors to address the remaining issues, and make the manuscript smoother to follow. Such an interesting topic merits publications and will be valuable for more accurate access on FSD.

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

The abstract talks more about the background, instead of the solid achievements in the present study. I suggest a shorter background, and more results of the present study should be presented.

Thanks for the suggestion! We have revised the abstract.

1. Is there any relationship between the airborne data in 2.1 and the satellite data in 2.2? Or both of them are employed here just to test the effect of the new method on different kind of imagery.

There is no relationship between the airborne data and satellite data.

The local-scale airborne MIZ data were used to train and test DL models. The satellite data were additional data only used to test the generalization ability of the DL models from local to global scale, and they were never encountered by the DL models during the training.

2. Lines 210-215. “U-Net++ with the depth of 5 achieved the best floe instance segmentation”, is this a result of “experiments to compare the performance of U-Net++ with other SoA semantic segmentation architectures”? I mean if you have known U-Net++ is the best among all, why do you compare them again? And for the other methods such as ResUNet, ResUNet++, additional explanations should be added here to tell the difference between them.

Yes, it is the result of the comparison among different DL models.

It is common to show performance comparisons between different DL models if using DL method. We have moved this model comparisons to the appendix and add diagrams of the models.

3. There are two fig10e. And for figures 9-10, the difference between these results are very difficult to distinguish if no additional notations such as in fig11e are presented.

Thanks for the suggestion! We have added notations to these figures and correct the wrong figure numbering.

4. It is a little difficult for me to follow the contents in sections 4 and 5. A possible reason is that so many names of processing methods are presented here, and also two kinds imagery are included as examples to show the effect of these methods. I was not told why airborne data were employed here but satellite imagery were employed there. Thus the main improvement of the present study are submerged by these information.

We apologize for the confusion. We have adjusted the structure of the manuscript and Section ‘‘Experimental results’’ is moved to the appendix.

5. There are some very interesting results in section 6, for the variations in the power-law exponent, can you give some more explanations on them? Otherwise, it is not necessary to present so many pictures as example without any discission.

It is common to use some typical images to show the effect of the method. Therefore, we chose these floe images with low to high ice concentrations to present segmentation results made by the DL-based methods.

The determination of the ice floe size distribution is a further application after the ice floe segmentation, and it is also part of our ongoing project. We thank for the suggestion, and we have given a brief explanation on it in the revised manuscript.

6. There is a so quick stop in the conclusion section, can you give some evaluations on the limitations of the present study?

Apologies for the short conclusion and thanks for the valuable suggestion! We have revised the Conclusion section and discuss the limitations of the present method in the Conclusion section.

Dear TC editor and authors of the revised manuscript egusphere-2023-295,

The manuscript topic is interesting and ice floe information is useful for multiple purposes. The other two reviewers have already provided good comments to improve the manuscript. I hope my comments will complement the comments of the other reviewers.

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

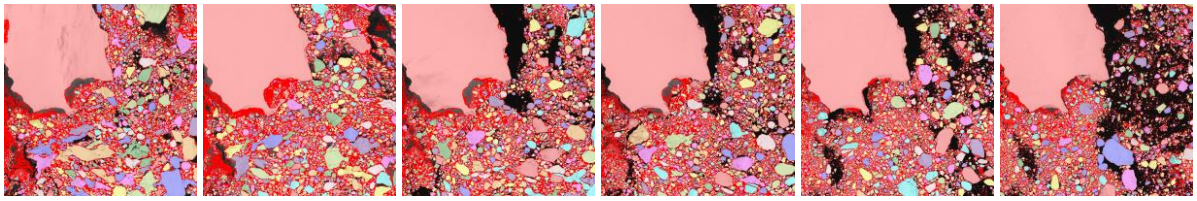
As suggested by another reviewer the abstract needs to be updated to include more on the applied method and results provided by the method instead of general level information. The abstract also begins with FSD and in the manuscript FSD is in the section named as "Case study:FSD" and FSD does not appear in the manuscript title. I suggest at least to remove "Case study:" from the section 6 title, also consider including FSD in the manuscript title. FSD is also significantly present in the introduction section.

Thanks for the comments. We have revised the abstract, reorganized manuscript structure, and included ''floe size distribution'' in manuscript title.

The amount of data is very restricted, Why only four Sentinel-2 images have been used? There exist a lot of Sentinel-2 data. It should be emphasized that with such limited data sets this is a case study and the results can possibly not be generalized.

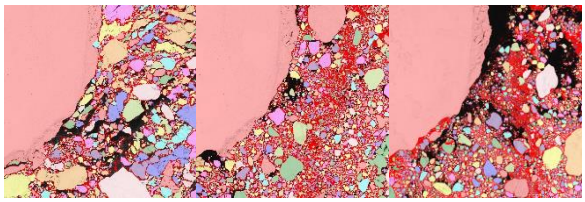
Our approach has been applied to extract the ice floes spatially and temporally from Sentinel-2 images, e.g., below show the floe segmentation results for S2-1 region from April to June in 2021 and other years.

2021



2020

2019



It is not practical to present so many images in the paper, and instead it is common to use some typical images to demonstrate segmentation performance of the methods. Therefore, in addition to four airborne MIZ images, we also used these four typical Sentinel-2 floe images with ice concentration from low to high to show the floe segmentation results. More examples of NetCDF files of Sentinel-2 images segmented for ice floes and containing both

georeferenced gridded and vector polygon data can be downloaded from https://thredds.met.no/thredds/catalog/digitalseaice/catalog.html. These contain ancillary data including a panchromatic quicklook of the original S2 image, Sen2Cor and FMask cloud masks, and a table of floe metrics including area, perimeter and long-axis lengths, and orientation.

Please note that the DL models were trained only on local-scale airborne MIZ images, and they did not encounter global-scale satellite data during the training. The Sentinel-2 images were extra data used to investigate the generalization abilities of the trained DL models from local-scale MIZ images to global-scale images (please see our response to reviewer 1's comment ''Line 98'').

Please note also that, how many individual floes (especially those that tightly connected with many other floes) are successfully identified from an image without being over- or under-segmented is also an important measure of ice floe segmentation method.

A proper cloud mask is required to be able to automatically segment ice floes. As "manual-free" is mentioned in the manuscript title I think cloud masking should be discussed in the manuscript. Does there exist automated methods for reliable cloud masking or at least excluding images with clouds? Give references of possible cloud masking approaches or suggestions for improved automated cloud masking. Could "manual-free" in the title be "automated" instead?

There is an option to choose cloud coverage when downloading Sentinel-2 data from the Copernicus Open Access Hub. So it can exclude images with clouds.

In addition, we have applied and compared the existing methods, Sen2Cor (via ESA Snap software or running on linux terminal) and Fmask, to mask clouds in Sentinel-2 images by means of cloud masking and classification. For us, Fmask works better than Sen2Cor. Please see below for reference.

Thanks for the comment, we have added a brief discussion on cloud issue in the revised manuscript, and revised manuscript title.

=====

Cloud masking for Sentinel-2 floe image segmentation

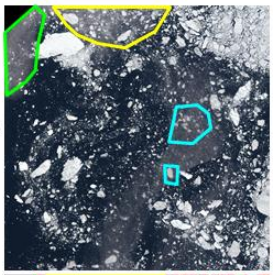
Thin Cloud

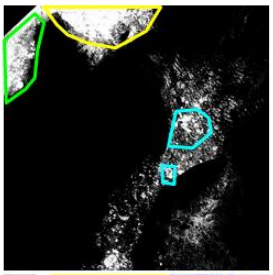
- Cloud mask: some floes may be mistaken as cloud

- Segmentation: some thin cloud pixels may be classified as ice edges (leading to over-estimation of SIC)

- Some floes under thin cloud can be identified by the model, but they may be masked out by the cloud mask

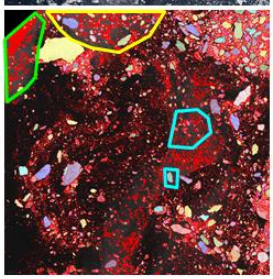
S2 image

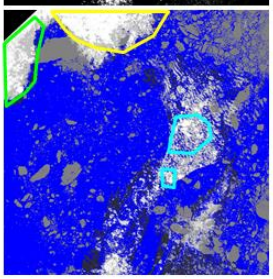




Cloud mask (sen2cor)

Segmentation





Classification (sen2cor)

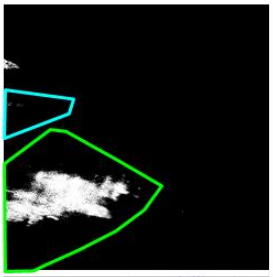
Thick Cloud

- Cloud mask: some cloud regions may not be detected

- Segmentation: it is hard to find floes and floe boundaries in and around the regions covered by thick cloud

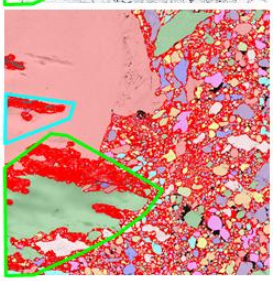
Suggestion:
The detected floes adjacent to cloud regions will not be considered when determining FSD, since they are likely to be incorrectly segmented

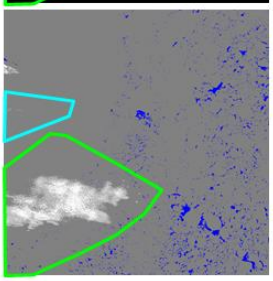
S2 image



Cloud mask (sen2cor)

Segmentation

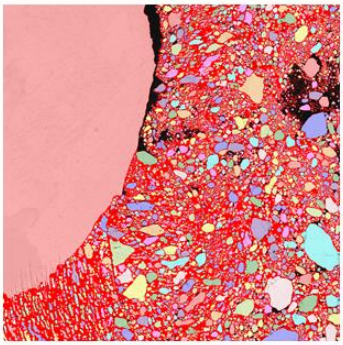




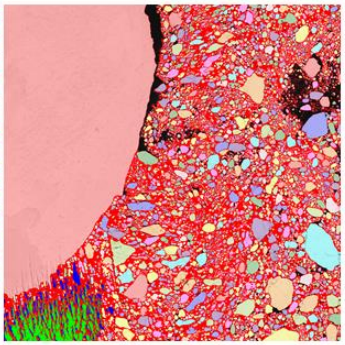
Classification (sen2cor)

S2 image

Before cloud masking

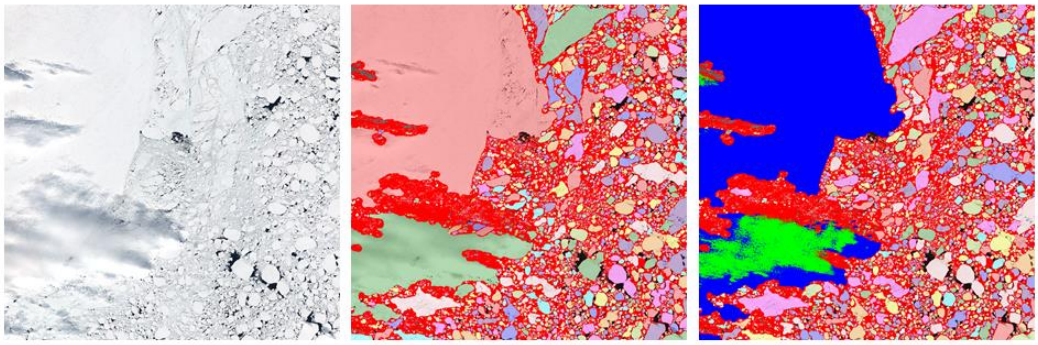


After cloud masking



Green: clouds, detected by sen2cor
Blue: floes adjacent to the detected clouds

S2 image | Before cloud masking | After cloud masking

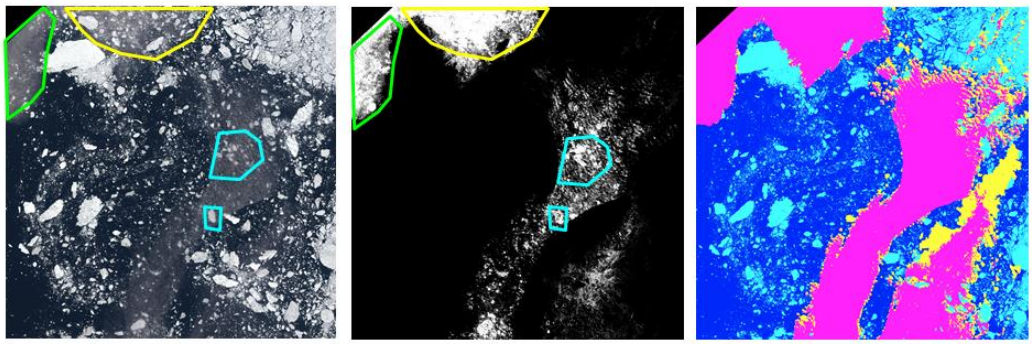


Green: clouds, detected by sen2cor
Blue: floes adjacent to the detected clouds

The current model doesn't consider the category of cloud/cloud shade, some cloud/cloud shade pixels are mistaken for ice edges.
Also due to the insufficient detection of cloud by the sen2cor module, it's hard to correct the this type of error.

Sen2Cor vs Fmask
Via cloud mask - thin cloud

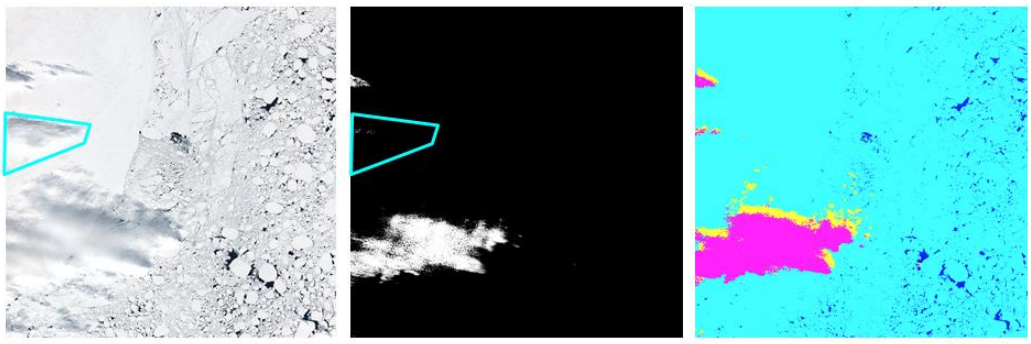
S2 image | sen2cor | Fmask



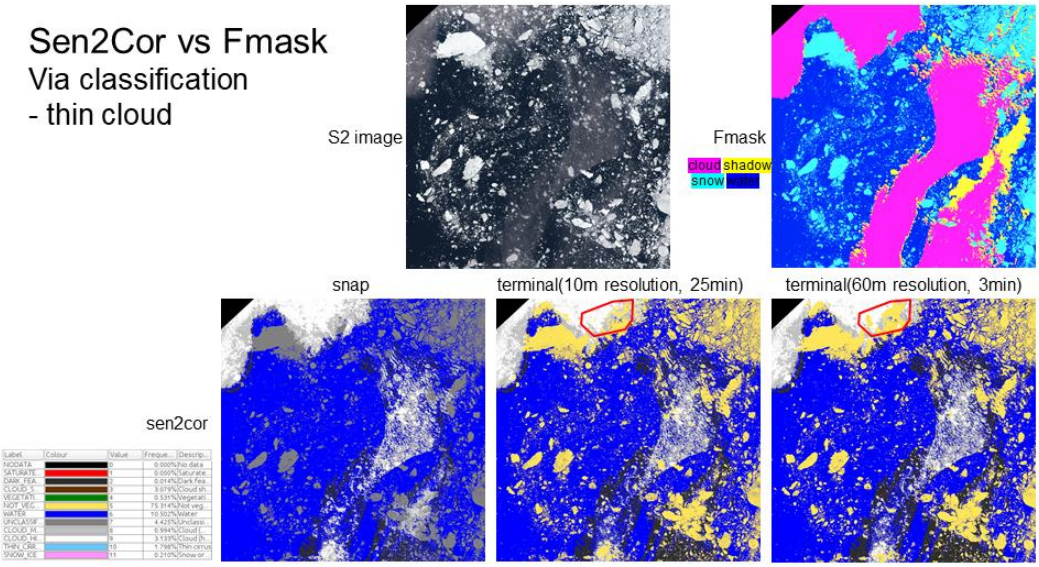
cloud shadow snow water

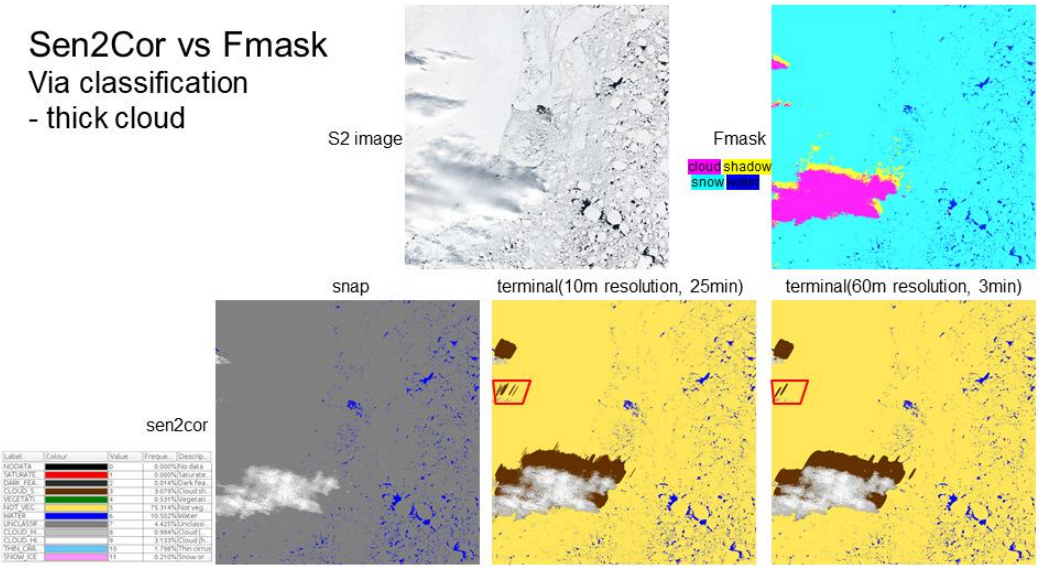
Sen2Cor vs Fmask
Via cloud mask - thick cloud

S2 image | sen2cor | Fmask



cloud shadow snow water





The results and discussion have now been presented in the same section. I suggest to make a separate “Discussion” section or rather combine “Discussion” with the “Conclusions” section which is very short now.

Thanks for the suggestion, we have combined the discussion with the conclusions.

P2 L36 “Copernicus”: give a reference

The link: https://scihub.copernicus.eu has been added there.

P3 Dataset → Airborne data: What is the number of airborne images used in this study? What is “a large amount”? Rather give numbers. Were the images used as long strips, mosaics, or single shots?

The total number of airborne images we got was 254, of which 52 were selected to be labelled for this study.

The airborne image data were used as single shots.

P3 Table 1: Give flight altitude(s) and surface area covered, or their ranges, for the images used. These could possibly be included in the table.

We apologize that we cannot provide the information you suggested.

Neither of the authors of this manuscript was on board the expedition 8 years ago. The airborne images were kindly provided from other research group at a university. Table 1 and the average resolution for a pixel were the only information we got about the airborne images.

P3 L83 “Cophub”: Give reference (URL).

The URL was given in the 1st ref. It has been moved to the end of ‘‘Copernicus Open Access Hub’’.

P4 Table 2: Include location information and covered area in the table, e.g. by given center latitude and longitude (and covered area e.g. in km2).

Thanks for the comment. We have included these information in the table.

P6 L110-112: Hypothesis of improvement by widening the boundaries would require some evidence. Would it be possible to show test results with a small set of imagery and some numeric evidence based on these tests?

We have added an appendix section to the revised manuscript to demonstrate the effectiveness of widening the ice floe boundary in improving model performance. Please see Appendix B for details.

P8 S3.2. Deep learning model: Give at least a short description of U-Net++ giving best results or a diagram of the network. Now this subsection is very short and it is essential for the study.

We apologize for the lack of descriptions about the models, and we have added description and diagram of the models in the revised manuscript.

P8 Post-processing: Applying morphological opening and closing seems a bit heuristic to me. Are there any references or if not would it be possible to demonstrate the benefits of the morphological processing? What is the shape and size of the morphological operator (often a disk is used)? Could this step be included in the ML algorithm somehow, i.e. could the NN learn the post-processing?

Please see (Banfield, 1991; Banfield 30 and Raftery, 1992; Soh et al., 1998; Steer et al., 2008; Wang et al., 2016) for reference. We also mentioned their work in the Introduction section:

‘‘Morphological operations can be used with different improvements to determine individual ice floes, but the methods operate directly on binarized floe images and thus cannot separate out the floes that had no or few gaps with any surrounding floes after binarization (Banfield, 1991; Banfield 30 and Raftery, 1992; Soh et al., 1998; Steer et al., 2008; Wang et al., 2016)’’

Due to the problems with the morphological operations described above, we used the morphological operations in a post-processing step to refine the floe segmentation. Fig. 5 and Fig. 6 in the manuscript can demonstrate the benefit of the processing. A disk-shaped structuring element with a radius of 4 pixels was used in the morphological operations. The disk shape was chosen because it is non-directional and can handle ice floes more uniformly than other shapes without being aware of floe’s irregular shape and orientation.

The next task after the work presented in the manuscript is to use the trained DL model together with the post-processing to label more ice floe images and create more dataset, and then train more robust DL models for ice floe segmentation, as we mentioned in the Conclusion section: ‘‘it still can be utilised as a ''higher version'' of ''annotation tool'' and produce more ''ground truth'' from a wide variety of ice image data sources, contributing to the establishment of datasets suitable for ice floe segmentation tasks, as well as further training more robust DL models for obtaining more accurate ice parameters from images.’’

P9 Training: GPU memory is referred on line 170 and this information is then given later on page 10. The hardware (and software) used should be given before it is referred. The used HW and SW could e.g. be included in the dataset section and changing the title to something like “Datasets and computational resources”. Also include the used SW with reference in the same section. Also mention there that all the execution times given later are given for this specific configuration.

Thanks for the suggestion! We have adjusted the structure of the manuscript.

P10 L174: Does this distribution of classes correspond to their distribution in general? Then it can be used in training. What happens if the distribution of classes is balanced (33% of samples for each class) for the training? Does balancing degrade the classification?

The distribution of classes depends on the extent of sea ice coverage and the amount of ice floes in the image. The higher the ice concentration in the image, the higher the proportion

of ‘‘ice’’ class (the lower the proportion of ‘‘water’’ class); the more ice floes, the higher the proportion of ‘‘floe boundary’’ class. MIZ images generally have higher proportion of ‘‘floe boundary’’ class than other ice floe images.

Balanced classes will not degrade the classification. Instead, it makes model easier to train because it helps the model learn the features of each class equally.

In ice floe segmentation, ‘‘floe boundary’’ is a hard-to-train class because: 1) the pixel intensity of floe boundaries, especially the boundaries between connected floes, is usually similar to that of ice; 2) the proportion of floe boundaries in ice floe images is usually much smaller than other two classes of "ice" and ‘‘water’’ (MIZ images are less affected by this issue). Therefore, it is necessary to increase the proportion of ice floe boundaries in the training data set, and using MIZ images as training images and widening floe boundaries can help with this.

Although the DL model was trained on MIZ images with a relatively high proportion of ice floe boundaries, it also has a good generalization ability to other ice floe images with low proportion of floe boundaries (e.g., the largest ice and water regions in Sentinel-2 images). It demonstrates the model trained on restricted datasets that can also generalize to wider datasets.

P10 L175: The number of test samples is not very large. What is the effect of reducing samples in training and validation data sets and increasing of the test data set? Are these numbers of samples selected based on some kind of performed tests?

Reducing the training and validation sets to increase the test set often leads to a decrease in the performance of DL models.

A common ratio between train, validation, and test is 80:10:10, and 70:15:15, 60:20:20, etc. are also practical ratios for splitting datasets.

Regarding the number of test samples, please see our response to your previous general comment ‘‘The amount of data is very restricted, …’’.

P11 Section 5: Would be good to have some kind of related introductory text under “5 Experimental results and discussions” and “5.1. DL model evaluation”, now they are empty.

We apologize for the lack of descriptions, and we have made some changes to the manuscript.

P13 Section 5.1.3. Inference time: Could this be “Segmentation time” instead, it would be more informative. The HW (and SW) used for segmentation could be given already in an earlier section, e.g. jointly with the introduction of data sets.

Thanks for the suggestion and we have changed them.

P15-15 Figure 10: Fig. 10 is nor in two parts and in two pages. Would it be possible to compress a little and make it to fot on one page?

We have put the figures on one page.

P19 “6 Case study: floe size distribution”: I recommend to drop “Case study:” because FSD is the essential parameter to be estimated by the method and it is also in essential role in the abstract and introduction and the whole manuscript is actually a case study because the datasets are quite limited.

Floe size distribution: Now FSD is estimated in two different resolutions (airborne and satellite data). It would be interesting to see how well FSD can be extrapolated from resolution to another (both from larger to smaller and smaller to larger) based on a fit distribution model. This would be very valuable information and this theme could be included in the discussion section.

Thanks for the suggestion! Multi-scale FSD is actually part of our ongoing project. We have given a brief explanation on FSDs in the revised manuscript.

P24 “Conclusions”: This section is very short. Possibly it could be combined with a discussion section. Here could also be some conclusions on how close to automated FSD estimation the proposed method is? Could it be used for operational monitoring and what will still be required before possible. At least cloud masking should be discussed and also the annual period of possible operation (lighting conditions, what is the fraction of cloudless time in suitable lighting conditions in different sea ice covered areas). What are the ways forward in automated ice floe analysis?

We apologize for the short conclusion and thanks for the valuable suggestion! We have revised the Conclusion section.