**Response to the Interactive comment on "Out-of-the-box calving front detection method using deep learning" by Oskar Herrmann er al.**

First of all, we want to thank the reviewers for their constructive comments on our manuscript. The following pages contain a list of reviewer comments followed by our replies. The comments are sequentially numbered and associated with the corresponding reviewer. The document contains active cross-references between the replies and the modifications in the main text. The black label references the original reviewer's comment within the manuscript. For instance, the reference **1.3**, refers to the response from the third comment of the first reviewer, the reference **A2/C2/D2 (3.5)**, which is typeset in the manuscript margin, refers to the second addition/change/deletion stemming from the fifth comment of the third reviewer.

### Comment of the Editor

*Some more detailed remarks from my side (that can be addressed during the revision process) are that when a new method is presented I think it is good practice to benchmark it to (some of the) existing methods. Otherwise, it might be difficult for the reader to effectively assess the (dis)advantage of the different methods and/or choose the optimal method for calving front detection.*

We sincerely thank the editor for the concern. A comparison between existing methods is highly needed. However, a problem arises: to compare the performance of different deep learning models, they need to be trained on the same data, tested on the same test set, and the same metric needs to be used for evaluation. Existing models have been trained and tested on different datasets (partly also different metrics were used). Hence, if we simply compare the results, we do not know whether the difference arises from the better-adapted model architecture, from a more representative train set, or from an easier test set. For this reason, Gourmelon et al. (2022) published a so-called benchmark dataset that is used for comparing the performance of different deep learning models. Our method is trained and tested on this benchmark and compared to the baseline (also presented in Gourmelon et al. (2022)) using the introduced metric. Re-training and testing other existing calving front extraction models on the benchmark is out of the scope of this study. We are, however, preparing such an inter-comparison project at the moment and hope to submit it soon.

### Comments of the 1st Reviewer:

**1.1** *So far, the results & discussion section is a bit difficult to follow and very short. Think about providing sub-headers and providing a more in-depth discussion.*

- Thank you for your suggestion. We added sub-headers (additions 10 and 11) and will include Fig. 11, which gives insights into the influence of resolution on the calving front prediction (comment 1.3). We added a more in-depth discussion on the seasonal impact (changes 13 and 14).

**1.2** *Addressing the seasonal effects on accuracy is very interesting. But a more sophisticated analysis would be required. In Figure 10 it looks like that the sensor type influences the accuracy much more than the season. Therefore, it would be interesting to know the sensor type for the scenes considered for winter and summer. Especially for Columbia glacier as only TSX and S1 scenes were available and for each sensor the MDE is very different.*

- Thank you for raising this interesting discussion. We include Table A4 in the Appendix and refer to it in addition 16. The Table shows the Mean Distance Error (MDE) grouped by satellite and seasons. The influence of the sensor on the MDE can also be seen in Fig. 12. Comparing the rows Mapple_TDX with Mapple_S1 and Columbia_TDX with Columbia_S1 gives the differences introduced by sensors without the influence of the test site. As you state, the sensor type influences the MDE much more than the season.

**1.3** *Provide an analysis on the effect of different spatial resolutions on the accuracy of the front extraction. Also introduce the different spatial resolutions and polarizations of all sensors as not every reader might be familiar with that.*

▸ Thank you for raising this point. We now introduce the different spatial resolutions in addition 4. We provide Fig. 11 and add analysis to the manuscript (addition 17). We added the resolution of each satellite in the caption. Unfortunately, the used benchmark dataset does not provide information about the polarization of the sensor for each image.

**1.4** *The conclusion is a bit unstructured and could be much stronger. Emphasize the identified best practice approach and give a better outlook for your study and applications in glaciology.*

▸ Thank you very much for your suggestions. We revised the conclusion by emphasizing the best practice approach and added a better outlook of the applications for the cryosphere community (change 17 and additions 22, 24 and 25)

**1.5** *L16: The GitLab repository is only accessible for people affiliated to the Friedrich-AlexanderUniversity. Please publish your scripts on a platform easily accessible for everyone.*

▸ Thank you for raising this issue. We uploaded the code on GitHub (https://github.com/ho11laqe/nnUNet_calvingfront_detection) and HuggingFace (https://huggingface.co/spaces/ho11laqe/nnUNet_calvingfront_detection). HuggingFace provides computing resources to execute programs directly via the web portal. This should be easily accessible to everyone (changes 1 and 18).

**1.6** *L45: "Most of these methods use optical imagery": Three out of four cited references here use SAR imagery. Please provide proper reference for your statement.*

▸ Thank you for pinpointing this inaccuracy. We changed the statement (changes 2 and 3 and addition 1).

**1.7** *L66: What means "visual preparation". I would assume a benchmark dataset can be used right away.*

▸ Thank you for pointing out this confusion. The benchmark can be used right away. With the term "visual preparation", we meant visualizations that provide an overview of the temporal and spatial distribution of the dataset. To be more precise, the term "visual preparation" refers to Fig. 4, 5, and Appendix C. We clarified this phrase in the revised version of the paper (change 4).

**1.8** *L76: Please do not only state that many methods are based on the U-Net from Ronneberger but also provide references for these studies/methods and mention that they mostly used modified versions of the original U-Net (e.g. Loebel et al. 2022, Mohajerani et al. 2019, Zhang et al. 2019 etc.)*

▸ Thank you for this suggestion. We added the references to the revised version of the paper and mentioned that they mostly used modified versions of the original U-Net (addition 2).

**1.9** *L98: In the study of Heidler et al. 2021 the digital elevation model was used during training for different experiments. A further development of the HED-UNet for a circum-Antarctic approach (Baumhoer et al. 2023) uses indeed a DEM in the post-processing.*

▸ Thank you for pointing out this inaccuracy. We edited this segment of the text and added the article from Baumhoer et al. (change 5 and addition 3).

**1.10** *L145ff: If the network itself decides on batch and patch size I wonder how big the effect on the accuracy is. For example, a smaller patch size would provide less spatial context and probably decrease the accuracy. Did you do any experiments on the effect of batch and patch size on the model accuracy? Furthermore, please state on what infrastructure (GPU type, RAM etc.) your model was trained and what patch and batch size was finally used.*

▸ This is an interesting point. We did not evaluate the effect of patch and batch size on the model. Our goal was to evaluate the nnU-Net's ability to segment the Synthetic Aperture Radar (SAR) dataset without modifications by us. The authors of the nnU-Net also suspect that a large patch size has a positive effect on accuracy. The nnU-Net makes the most of the available GPU memory for best performance. For reproducibility, we added a subsection about the experimental setup in the manuscript (addition 8). We trained the nnU-Net on an NVIDIA RTX 3080 with 12GB memory. This allowed a patch size of 1280x1024 for the experiments that had only one label. The experiments with more than one label have a patch size of 1024x896. For all experiments, the batch size is 2.

**1.11** *L225: Maybe add a sentence why you propose that the fused label approach is the best even though the boundary label approach has better accuracies.*

▸ Thank you for pointing out the confusing sentences. The mean accuracy of the boundary experiment might be better after our training run, but the difference is insignificant. We evaluated a T-test in table A1 to reveal significant differences. Therefore we choose this training approach because it uses the smallest amount of parameters and does not need a change of architecture as in the Multi-Task-Learning (MTL) experiments. We edited the sentence to make clear why we chose the fused label approach (change 12).

**1.12** *L245: Dry snow is penetrated by active microwave sensors and the ground can be the major source of the backscatter signal (depending on snowpack thickness and wavelength). Wet snow can be detected by SAR sensors. Hence, is snow cover really the reason for less accurate front predictions in winter? Additionally, I would assume that surface melt in summer and sea ice in winter make the front prediction more difficult. Did you investigate this further?*

▸ We understand the reviewer's concerns. Not only the snow cover and its properties, also the surface melt, and sea ice make the front prediction more difficult. In particular, the ice melange is certainly causing some issues when detecting the calving front. Even for an experienced human mapper, it can be very challenging to distinguish between the actual glacier tongue and the ice melange. We adjusted the wording of the respective section accordingly (changes 13 and 14).

**1.13** *L228: Why did you decide to use the ensemble prediction results for the final evaluation instead of doing this evaluation for the most feasible method you identified prior to that (fused label approach)? Wouldn't it be more likely for users to use this 'best' method instead of the ensemble approach requiring much more computational power and effort?*

▸ Thank you for pointing out the confusing explanation. The final evaluation does concern only the most feasible method. The ensemble consists of the five models that are trained during the five-fold cross-validation. Every model of this ensemble is trained on the fused label, but each one is trained on a different train-validation split of the training set. We added information to the manuscript to clarify this confusion (additions 12, 13 and 15).

**1.14** *L248: You can not directly compare the accuracies for both sensors. For ERS, Envisat and Palsar you only have data for Mapple glacier which seems to have a front that is better captured by the nnUNet. The accuracy comparison for different sensors is very interesting especially regarding spatial resolution and polarization but this would require a more sophisticated comparison.*

▸ Thank you for pointing this out. We clarified that ERS, ENVISAT, and PALSAR only capture Mapple Glacier (change 15 and addition 18). Moreover, we split the satellites into glaciers as well in Fig. 12. We refer to our answer on the comment 1.3 where we group the results by resolution. Unfortunately, the dataset does not provide information about the polarisation of the sensor.

**1.15** *L250: I would recommend to investigate further why the results are worse for Sentinel-1. Maybe provide some plots in the appendix on these in-accurate results plotted over the Sentinel-1 image to see the sea ice conditions at that time.*

▸ Thank you for this suggestion. It seems like the nnU-Net is more vulnerable to being distracted by ice melange in the images of Columbia taken by Sentinel-1 (S1), but there are also predictions where the calving front is in the middle of the ice-free ocean. We suspect that the complex Columbia calving front requires high resolution for accurate detection, due to the lower MDE of the TDX images of Columbia or more training samples of complex calving fronts with low resolution. We stick to our resilient conclusion that our method performs insufficient on the group of images of Columbia captured by S1 (addition 20). Further experiments that could verify our hypothesis that for the Columbia Glacier, a higher resolution than that of S1 is needed are out of the scope of this manuscript. However, we provide all predictions of the test set on Zenodo (https://zenodo.org/record/7915740#.ZFtS3dIzZhE). We added the link to the predictions in the manuscript (addition 21).

**1.16** *Figure 10: Please properly label the X-axis for the logarithmic scale. Consider splitting the figure in two or insert a little gap between the seasonal analysis and the comparison of accuracy for different satellite sensors. For ERS, Envisat and Palsar you need to mention that this data is for Mapple only.*

▸ Thank you very much for the helpful comments. We split the results plot into two. One contains the glacier and season comparison and the other contains the satellite comparison. We changed the labels of the X-axis and stated that ERS, Envisat and PALSAR only capture the Mapple glacier (Fig. 10, Fig. 12).

**1.17** *Figure 11c: The yellow front seems to be a circle. One part of the prediction fits very well whereas the other is completely off. How is this considered in the accuracy estimation?*

▸ This raises an interesting point. For each pixel in the predicted front, the distance to the closest pixel in the labelled front is computed, and for every pixel in the labelled front, the distance to the closest pixel in the predicted front is computed. The mean of all the distances is taken as the MDE. The MDE of this particular calving front prediction is $4221 \pm 5026\,\mathrm{m}$ (addition 19). The metric does not take into account geometric impossibilities like a calving front forming a circle. However, one side of a circle will still be far from the ground truth, and therefore, it will still negatively impact the MDE. We stick to this metric to enable comparability with the baseline presented by Gourmelon et al. (2022).

**1.18** *L269: Rather “increase” model performance I guess?! To which sensor do you refer when talk about low resolution imagery?*

▸ Thank you for this suggestion. We changed the word "benefits" to "increase" (change 16). We mean the S1 sensor with $20\,\mathrm{m}$ per pixel resolution (addition 23).

**Comments of the 2nd Reviewer:**

**2.1** *The author applies a nnU-Net and improves the error from 753 ± 76 m by Gourmelon et al. (2022) to 541 ± 84 m. Still, 541 meters is a relatively large error for terminus delineation, compared with previous studies that have errors ranging from 33 to 108 meters (Zhang et al., 2019; Cheng et al., 2021; Baumhoer et al., 2019). Given the relatively large margin of error in the terminus products produced by this method, it might be challenging to conduct scientific research that relies on them. Also, providing more termini that go beyond the benchmark dataset might be more beneficial to the glaciology community, but I understand it might be out of the scope of this study.*

▸ We sincerely thank the reviewer for her/his concern. First of all, we would like to state that this manuscript is a methodology paper, which aims to improve the method used to extract calving front positions from SAR imagery (semi-)automatically. Therefore, as the reviewer already kindly states, it is out of the scope of this study to provide additional termini outside the benchmark dataset. The reviewer also correctly states that researchers should not directly rely on terminus products provided by our method. Our method should be used semi-automatically, like, e.g., CALFIN (Cheng et al., 2021), where a manual verification of each output is done with the original image (only this manual verification leads to a mean distance error < 100 m). We do not claim that our method perfectly extracts all calving fronts, and we now state more clearly in our conclusion that the output needs manual verification (see addition 22).

However, we object the direct comparison of our study results with the result from previous studies. To correctly assess the performance difference between two deep learning models, the models need to be trained on the same training data, tested on the same test set and evaluated with the same metrics. Otherwise, it is not possible to discern where the performance difference stems from.

Here we would like to highlight some reasons in particular that prohibit a comparison with previous studies:

(1) In some studies, an in-sample test set is used (e.g., Zhang et al. (2019), Hartmann et al. (2021) and partly also Cheng et al. (2021)), i.e., the test set contains images of glaciers that are also included in the train-set, but at different time points. A test on an in-sample test set is easier, as the model does not need to generalize to a new environment.

(2) In comparison to our test set, some other test sets are not as challenging (e.g., Zhang et al. (2021) - Helheim Glacier features only one calving front with a simple geometry) or rather small (Baumhoer et al. (2019) - 8 samples all from the same month).

(3) Other test sets, in turn, have only high-resolution imagery (e.g., Zhang et al. (2019)), which we at least hypothesize helps the automatic calving front extraction with deep learning methods.

(4) Baumhoer et al. (2019) delineate the complete ice shelf/ice sheet coastline, not a glacier calving front bounded by bedrock. Hence, a direct comparison can not be drawn.

(5) Our test set, in contrast to others (e.g., Baumhoer et al. (2019) and Zhang et al. (2019)), features multiple sensors introducing more variability and, therefore, making it more challenging for deep learning models.

(6) The comparison to the averaged performance value of Cheng et al. (2021) lacks in validity two-fold: 1. The average performance is in big parts calculated over results from optical imagery. CALFIN's results on subsets of the test set that only include SAR imagery are 115.24 m on Zhang et al. (2019)'s test set and 330.63 m on Baumhoer et al. (2019)'s test set. 2. The performance metrics are calculated after manual verification. Thus, outliers that would lead to a higher MDE are not included anymore.

(7) In general, two metric values are important: the MDE and the number of samples with no predicted front. A trade-off between the two has to be made. For difficult samples, is it desirable to predict a front with a high MDE, or is it better to predict no front at all? For example, Cheng et al. (2021) sort out unreliable predictions by confidence scores and by hand and, thus, receive a lower MDE but a higher number of samples with no predicted front (on the Zhang et al. (2019) test set $4 \in 12$ results are filtered out, while for our method only $5 \in 122$ predictions show no front).

To improve the comparability of deep learning models for calving front extraction from SAR imagery, Gourmelon et al. (2022) published a so-called benchmark dataset with a corresponding evaluation metric. We only compare to the baseline models provided in (Gourmelon et al., 2022) and improve over the baseline's performance. A re-training and re-evaluation of previous calving front extraction models is out of the scope of this study. However, we are preparing such an intercomparison study so that this study's method can be compared with previous studies in a rigorously correct manner.

**2.2** *One key advantage of using no new U-Net is that the framework can take the fingerprint of the dataset and automatically configures the framework (adjust the hyperparameters). However, it is not clear to me which hyperparameters are being adjusted automatically, how they are adjusted, and what is the final choice of these hyperparameters. If we adopt the hyperparameters from nnU-Net and apply them to standard U-Net, would there be no difference between nnU-Net and U-Net? If that is the case, I suppose it is important to clearly demonstrate how to automatically adjust the hyperparameters and how these newly derived hyperparameters improve the results.*

▸ Thank you for pointing out this incomplete description of the nnU-Net. A lot of hyperparameters are fixed by the authors of the nnU-Net. This includes the optimizer (Stochastic gradient descent (SGD)) with an initial learning rate of 0.01, a Nesterov momentum of 0.99, and a weight decay of 3e-5. One epoch is defined as 250 iterations. The batch and patch size of one iteration depends on the available GPU memory. In our case, that leads to a patch size of 1280x1024 for the experiments with only one label. The experiments with more than one label have a patch size of 1024x896. For all experiments, the batch size is 2. The number of layers of the neural network depends on the patch size. The authors introduce building blocks containing layers for the nnU-Net (see Fig. 1). Each block reduces the spatial dimensions. Blocks are added to the network until the feature maps have a shape of 4x4 pixels. The building blocks of the nnU-Net contain Instance Normalization, which the vanilla U-Net didn't, they use Leaky ReLU instead of ReLU as an activation function, and they use a stride of [2,2] instead of max-pooling to reduce spatial dimensions.

In addition, to the hyperparameters, the nnU-Net has sophisticated training techniques, that are not included in the vanilla U-Net. The nnU-Net uses Deep Supervision. This technique avoids vanishing gradients in deep neural networks (containing many layers). An additional output of each decoder block is compared to a down-scaled version of the target label. The error is multiplied by a weight and added to the final loss. The weight increases towards the final layer.

Besides the architectural changes, there are also changes in the training and inference procedure. In the work of Ronneberger et al. (2015), they already used data augmentation (shift, rotation, random elastic deformation, and gray value variations). The nnU-Net framework extends the list of augmentation techniques (see Table 1). Another technique for the inference procedure sets the vanilla U-Net and the nnU-Net framework apart. For more robust predictions, a sliding window with half-patch size overlap and Gaussian patch centre weighting is used for inference. Additionally, each patch is rotated three times (90°, 180°, 270°) and the predictions of each rotation are fused to a more robust prediction.

To summarize, the nnU-Net of Isensee et al. (2021) is a framework around the U-Net architecture. In addition to providing good default values for hyperparameters and rules to adapt hyperparameters to the dataset, it contains many established deep-learning techniques for training and inference that make predictions more robust.

We added a new subsection to the manuscript, which provides the hyperparameters of the nnU-Net generated and information about the experimental setup (addition 8). We made changes in the Background (change 7).

The improvement of the nnU-Net framework techniques over another version of the U-Net is given by the comparison to the work of Gourmelon et al. (2022) (see Fig. 9, Fig. 10, and Fig. 12).

**2.3** *While the author emphasizes that the method is an out-of-the-box application, it would be beneficial for the readers to have a clearer understanding of the steps taken by the author to ensure its out-of-the-box applicability. Providing further details on this aspect could enhance the credibility of the method's out-of-the-box applicability claim and contribute to its broader adoption in the research community.*

▸ Thank you very much for the helpful comment. This was actually a misunderstanding. With "out-of-the-box" we initially did not want to imply that a reader can use our pre-trained model to detect calving fronts in their satellite images with little effort. Instead, we meant that we use the unmodified (out-of-the-box) nnU-Net on the dataset of Gourmelon et al. (2022). We now see that this was misleading. Regardless of the interpretation of the title, we want to make our method easily accessible - to make our trained model *itself* useable out-of-the-box. We provide two ways to use our pre-trained

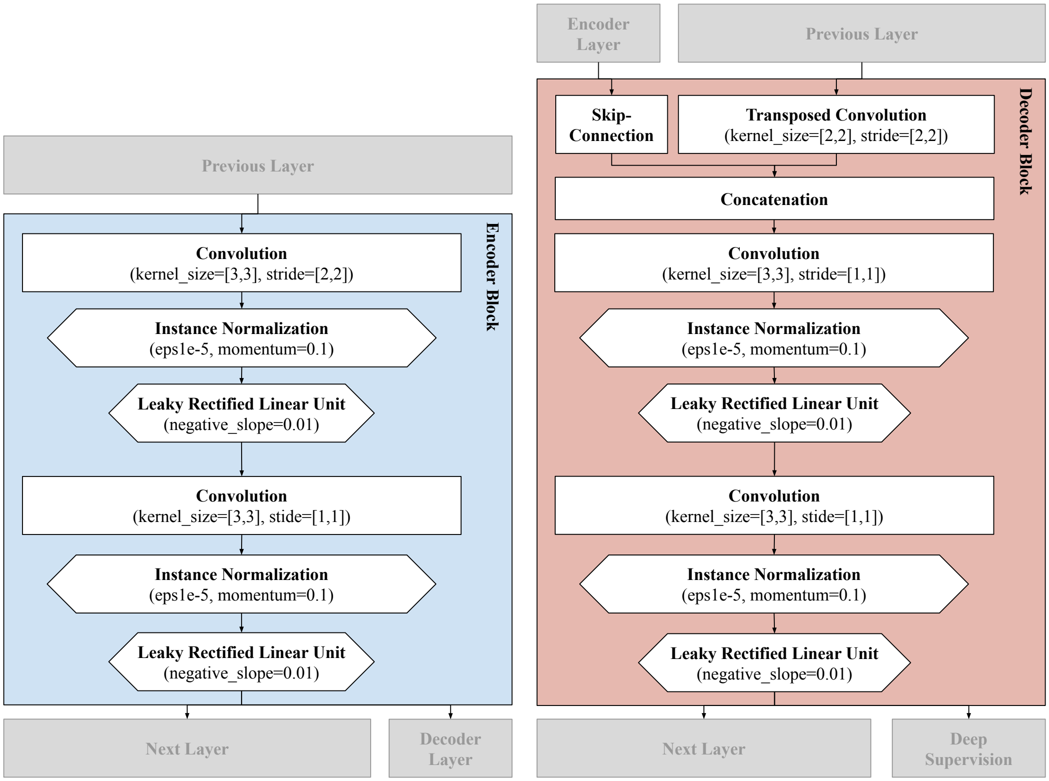


**Figure 1.** Illustration of the encoder and decoder blocks that make up the architecture of the nnU-Net. The encoder and the decoder contain multiple blocks.

| Augmentation | Parameters | Probability |
|---|---|---|
| Elastic Deformation | | 0% |
| Rotation | x=[-3.14, 3.14] | 20% |
| Scaling | factor=[0.7, 1.4] | 20% |
| Random Crop | | 0% |
| Gaussian Noise | add noise | 10% |
| Gaussian Blur | add Blur sigma=[0.5, 1] | 20% |
| Brightness Multiplication | range=[0.75, 1.25] | 15% |
| Contrast Augmentation | | 15% |
| Low Resolution | zoom=[0.5, 1] | 25% |
| Gamma | gamma=[0.7, 1.5], | 30% |
| Mirroring | x- and y-axis | 50% |

**Table 1.** List of augmentations used by the nnU-Net Framework. The table includes the parameters of the augmentations and the probabilities of their usage.

nnU-Net for calving front detection. One is via GitHub (https://github.com/ho11laqe/nnUNet_calvingfront_detection) with a short instruction for installation and application on SAR images. An even simpler way is provided with HuggingFace (https://huggingface.co/spaces/ho11laqe/nnUNet_calvingfront_detection). The code is executed directly on the website, so the reader does not have to install any Python packages. The drawback is that the code runs much slower than on a PC with a GPU, so it is suited only for a small number of SAR images. We added the links to the according paragraphs in the manuscript (changes 1 and 18).

**2.4** *Line 16: The scripts are not publicly available. As the author emphasizes that the method is out-of-the-box, it would be more beneficial if the codes are publicly available.*

- Thank you for pointing this out. We uploaded the scripts to GitHub (https://github.com/ho11laqe/nnUNet_calvingfront_detection) and HuggingFace (https://huggingface.co/spaces/ho11laqe/nnUNet_calvingfront_detection) to make it publicly available (change 1).

**2.5** *Line 66: What is the visual preparation? Why is the visual preparation necessary as the dataset from Gourmelon et al. (2022) is a benchmark dataset?*

- Thank you for pointing out this inaccurate term. No visual preparations of the dataset are necessary. With the term "visual preparation", we meant visualizations that provide an overview of the temporal and spatial distribution of the dataset. To be more precise, the term "visual preparation" refers to Fig. 4, 5, and Appendix C. We clarified this phrase in the revised version of the paper (change 4).

**2.6** *Line 115: Are sentences about the labels here and the sentences on Line 50 repeated?*

- Thank you for pointing this out. The content of these lines does repeat. After considering shortening the sentence in the introduction, we keep the composition of the labels in both chapters because they occur in two separate chapters and transport important information to the reader.

**2.7** *Figure 3: For Crane, what is the black dots in the middle of the ocean?*

- Thank you for your interest. The black dots in the ocean zone of Crane are areas with no data available due to a small artefact at the former calving front of Crane Glacier in the DEM (Cook et al., 2012) used for the orthorectification of the SAR data. We added the information to Fig. 3.

**2.8** *Line 120: Are there any procedures to deal with the imbalance? It might be helpful to increase the weight of the positive pixel loss in the loss function.*

- The reviewer raises an excellent point. There is a great imbalance between front pixels and background pixels. One technique of the nnU-Net to reduce the class imbalance is foreground oversampling. It is ensured that (at least) one third of the patches for training are guaranteed to contain a foreground class. In our case, every batch contains two patches, from which one is forced to contain calving front pixels.

  Additionally, the Dice-loss is part of the final loss. The Dice-loss is defined in Eq. 1.

$$l_{dice}(y) = \frac{2|X \cap Y|}{|X| + |Y|} \tag{1}$$

Where $|X|$ represents the number of pixels assigned to one class in the prediction, and $|Y|$ is the number of pixels of a class in the target. $|X \cap Y|$ the number of pixels assigned to the same class in prediction and label. Because the number of pixels per class is included in the denominator and is computed in each mask separately, the loss is normalized so that every class is treated equally (Li et al., 2019a).

We added an explanation about how the class imbalance problem is faced in additions 5 and 6.

**2.9** *Figure 5: Please explain the abbreviations of the glacier names.*

- Thank you for your comment. We added an explanation of the abbreviations in the caption change 6 and in the caption of Fig. 5.

**2.10** *Section 4.1: Additional information about the method needs to be provided. It would be beneficial to include more details on the data fingerprint and how to use the fingerprint to determine the hyperparameters in this study. The following are some specific descriptions that is unclear to me:*
*What is the definition of the distribution of spacing?*
*How to use the distribution of spacing to determine the annotation, image resampling strategy, and target spacing?*
*Is the annotation equivalent to labeling the images? If so, how is that related to the distribution of the spacing?*
*What is the image resampling strategy that is used?*

- Thank you for this clarifying question. The distribution of spacing refers to three-dimensional (3D) medical data like Computer Tomography (CT), with different spacing between layers. Since the SAR have equal spacing, no re-sampling strategy is necessary. We now omit parts of this section since they are not relevant to our experiments (changes 8 and 9).

**2.11** *Is CT or z-score normalization used in this study? How to conduct the z-score normalization with image mean and standard deviation?*

- Thank you for the clarifying question. Yes, z-score normalization is used. The mean is subtracted from the pixel value and divided by the standard deviation. We clarify this in change 10.

**2.12** *Line 175: Please be more clear about the minor changes here.*

- Thank you for your comment. By minor changes, we mean changing the number of channels in the case of MTL. We clarified this in change 11.

**2.13** *Line 180: What is the post-processing step for the fifth experiment? How to combine the three types of output to get a single glacier terminus?*

- Thank you for pointing out the confusing description of the post-processing. Only the zone output is used to get the final front position. The other labels help the nnU-Net during training to adjust the weights so that they fit better to the application (addition 7).

**2.14** *Line 188: "To generate a prediction of the zones, pixels classified as front are assigned to the ocean, and the glacier zone is dilated once with a 7x7 kernel." Do the zones here represent the segmentation zone? If so, how does this lead to a glacier terminus? Similar to the fifth experiment, how to combine the two types of output to get a single glacier terminus?*

- Thank you for your comment. Yes, the zones in this sentence represent the segmentation zone. We added the post-processing to extract the zone segmentation and the front segmentation from the fused prediction. This way, we can compare it to the other MTL experiments. For the final evaluation, we omit the segmentation metrics (table A3) and focus on the MDE. The counterintuitive part is that we use the zone segmentations to get the front position instead of directly taking the front segmentation. There are predictions with a neighbouring glacier and ocean zone with no front prediction in between. We assign the ocean pixels that have neighbouring ocean pixels as front pixels. Therefore the front position that is extracted from the zone can differ from the front segmentation even in the fused label approach. We mentioned this in section 4.2, but we will add it also to the section Analysis of the fused label experiment (addition 14) and section Evaluation (addition 9).

**2.15** *Line 219: Why the errors of the STL approaches here are larger than the error of Gourmelon et al. (2022)?*

- This raises an interesting point. After adaptation, the nnU-Net has eight pooling (and un-pooling) steps, whereas Gourmelon et al. (2022) have only four pooling steps but introduce Atrous Spatial Pyramid Pooling (ASPP) in the bottleneck. One hypothesis is that ASPP works better than simply increasing the pooling steps (and, with that, the encoder blocks). However, the success of both pooling and ASPP also relies on the patch size, which is another unknown in the question of why the performance of the networks differ. There are many differences between the two models, and it is out of the scope of this study to test which subset of these many changes makes the difference. We can only give a more detailed comparison between the two methods by including Table A5.

# Out-of-the-box calving front detection method using deep learning

Oskar Herrmann[1], Nora Gourmelon[2], Thorsten Seehaus[1], Andreas Maier[2], Johannes J. Fürst[1], Matthias H. Braun[1], and Vincent Christlein[2]

[1]Institute of Geography, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
[2]Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

**Correspondence:** Oskar Herrmann (oskar.herrmann@fau.de)

**Abstract.** Glaciers across the globe react to the changing climate. Monitoring the transformation of glaciers is essential for projecting their contribution to global mean sea level (GMSL) rise. The delineation of glacier-calving fronts is an important part of the satellite-based monitoring process. This work presents a calving front extraction method based on the deep learning framework nnU-Net, which stands for no new U-Net. The framework automates the training of a popular neural network, called U-Net, designed for segmentation tasks. Our presented method marks the calving front in Synthetic Aperture Radar (SAR) images of glaciers. The images are taken by six different sensor systems. A benchmark dataset for calving front extraction is used for training and evaluation. The dataset contains two labels for each image. One label denotes a classic image segmentation into different zones (glacier, ocean, rock, and no information available). The other label marks the edge between the glacier and the ocean, i. e., the calving front. In this work, the nnU-Net is modified to predict both labels simultaneously. In the field of machine learning, the prediction of multiple labels is referred to as Multi-Task-Learning (MTL). The resulting predictions of both labels benefit from simultaneous optimization. For further testing of the capabilities of MTL, two different network architectures are compared, and an additional task, the segmentation of the glacier outline, is added to the training. In the end, we show that fusing the label of the calving front and the zone label is the most efficient way to optimize both tasks with no significant accuracy reduction compared to the MTL neural network architectures. The automatic detection of the calving front with a nnU-Net trained on fused labels improves from the baseline Mean Distance Error (MDE) of $753 \pm 76\,\mathrm{m}$ to $541 \pm 84\,\mathrm{m}$. The scripts for our experiments are published on ~~Gitlab (https://gitlab.cs.fau.de/ho11laqe/nnunet_glacer.git)~~ GitHub (https://github.com/ho11laqe/nnUNet_calvingfront_detection). An easy access version is published on Hugging Face (https://huggingface.co/spaces/ho11laqe/nnUNet_calvingfront_detection).

C1 (1.5, 2.4)

## 1 Introduction

Unlike the large majority of land-terminating glaciers, marine and lake-terminating (MALT) glaciers reach a water body at a low elevation. The contact surface is often referred to as the calving front. Their ice is lost by sub-marine melting or calving, i. e., ice that breaks off, gets de-connected, and starts to float freely in the form of icebergs or ice floes. Both processes, sub-marine melting and calving, determine the total frontal ablation, i. e., the mass loss at the calving front. Frontal ablation is often a dominant factor in the total mass budget of MALT glaciers (McNabb et al., 2015; Shepherd et al., 2018; Minowa et al., 2021). Besides its importance in the total glacier mass balance, the representation of processes controlling frontal ablation is currently a pressing task for numerical glacier models (Beer et al., 2021). Neglecting frontal ablation can introduce an important bias. Recinos et al. (2019) analyzed the impact on ice thickness reconstruction (based on mass conservation) in Alaska and reported an underestimation of 19 % on regional and up to 30 % on glacier scales. Various successful approaches exist to parameterize frontal ablation for individual glaciers. Still, the implementation in large-scale or global models is limited by the amount and quality of measurements to constrain the models (Recinos et al., 2019). Thus, large-scale measurements (ideally time series) of frontal ablation are demanded by the modelling community (Recinos et al., 2021). Driving forces of frontal ablation are, on the one hand, ice flux (higher flux, e. g., due to bed lubrication by meltwater can trigger calving events) and, on the other hand, marine factors such as ocean temperature, fjord bathymetry, and sea ice/ice mélange conditions (Carr et al., 2014; Straneo et al., 2013). For example, the persistence of the ice mélange in front of the glacier can stabilize the calving front and affect the glacier dynamics. In contrast, the breakup of the ice mélange can lead to increased calving and ice flow at the glacier terminus (Amundson et al., 2010; Kneib-Walter et al., 2021; Rott et al., 2020). Moreover, a significant frontal retreat can also indicate a retreat of the grounding zone (Friedl et al., 2018). The retreat of the grounding zone of a glacier with retrograde bedrock formation will lead to further grounding zone retreat, resulting in increased ice loss and destabilization of the glacier or ice stream (Robel et al., 2016). Thus, information on the temporal variability of the calving front position provides fundamental information on the state of the glacier or ice stream. Therefore, the glacier area has been defined as an Essential Climate Variable (ECV) product by the World Meteorological Organization (WMO). Calving front positions were usually manually mapped using different remote sensing imagery (Baumhoer et al., 2018). Only a few studies applied automatic or semi-automatic approaches. In polar regions, the ocean downstream of the glaciers is often covered by sea ice and calved off icebergs, forming the so-called ice mélange, making calving front delineation a challenging task, even when captured by hand. Deep learning approaches have shown high potential for carrying out such complex segmentation tasks, e. g., on medical imagery (Jang and Cho, 2019). In recent years, the application of deep learning techniques for glacier front detection started ~~(Zhang et al., 2019), (Baumhoer et al., 2019), (Cheng et al., 2021), (Hartmann et al., 2021).~~(Zhang et al., 2019; Cheng et al., 2021; Baumhoer et al., 2019; Hartmann et al., 2021; Mohajerani et al., 2019; Baumhoer et al., 2021; Zhang et al., 2021; Heidler et al., 2021; Marochov et al., 2020; Loebel et al., 2022). ~~Most of these methods use optical imagery, which has the drawback of being dependent on daytime and cloud coverage. Additionally these published methods are incomparable since the algorithms or the input data remained undisclosed.~~Calving fronts can be located in both optical and SAR imagery. In optical imagery, calving fronts are more easily distinguishable, whereas SAR imagery has a higher scene availability, as it is independent of daytime, season and cloud coverage (Baumhoer et al., 2018). A direct comparison between the results of existing deep learning-based calving front extraction studies is not possible, as the models have been trained on different data, tested on different test sets, and evaluated using slightly differing metrics.

C2 (1.6)

C3 (1.6)

The benchmark dataset published by Gourmelon et al. (2022) provides 681 SAR images of calving fronts. SAR imagery is independent of sunlight and cloud coverage, enabling continuous temporal coverage of the observation area, but compared to optical data, it has only one channel and has more speckle noise. For every SAR image, two labels are provided. One label provides four classes: ocean, glacier, rock, and no information available (e. g., radar shadow and layover areas, areas outside the swath). The other label marks the calving front with a one-pixel wide line. Based on the training set of the dataset, Gourmelon

A1 (1.6)

et al. (2022) train a modified U-Net for each label. One U-Net solves the task of glacier segmentation, and one detects the calving front. On the two test glaciers of the dataset, the segmentation model achieves an Intersection over Union (IoU) of $67.7 \pm 0.6$. By taking the boundary between ocean and glacier, they extract a prediction of the calving front from the zone prediction with a Mean Distance Error (MDE) of $753 \pm 76\,\mathrm{m}$. The model trained directly on the front label achieved an MDE of $887 \pm 189\,\mathrm{m}$.

In this work, we present a method that utilises both labels of the dataset instead of training separate models for each task (see Fig. 1). Multi-Task-Learning (MTL) is a technique for machine learning algorithms that uses one model to tackle multiple tasks. In most cases, the potentially larger dataset and higher information content lead to higher performance for the individual tasks (Bischke et al., 2019; He et al., 2021; Li et al., 2019b; Amyar et al., 2020; Chen et al., 2019).

Our method is based on the nnU-Net (no new U-Net) proposed by Isensee et al. (2021), which is an out-of-the-box framework for training the U-Net. The framework takes a fingerprint of the dataset and adjusts the hyperparameters accordingly. Therefore, the nnU-Net is a powerful tool that simplifies the application of deep learning algorithms. Its performance on segmentation of SAR data has not been tested, and the availability of two labels suggests the modification of the nnU-Net for MTL. As a baseline for our evaluation of the nnU-Net, we use the results of Gourmelon et al. (2022).

In particular, our contributions are as follows: ~~(1) Visual preparation~~(1) Visualization showing the temporal and spatial distribution of the dataset by Gourmelon et al. (2022). (2) Out-of-the-box application and evaluation of the nnU-Net for calving C4 (1.7, 2.5) front detection and zone segmentation. (3) Evaluation of two different U-Net architectures for MTL. (4) Test if an artificial third label improves the calving front detection. (5) Introduction of an efficient MTL approach that fuses the two labels of the dataset. (6) Analysis of the influence of season, glacier, and satellite on the performance of the model.

The paper is organized as follows: After presenting the related work in Sec. 2, we give an overview of the dataset (Gourmelon et al., 2022) in Sec. 3. Section 4 explains the method and our six experimental setups. Section 5 examines the results and analyses the influence of different properties of the satellite images. Section 6 summarises the work.

## 2 Related Work

Automated monitoring of glacier-covered areas is a growing research field. Recent glacier monitoring uses deep learning methods due to the increasing availability of satellite images and computing power. Many methods are based on the U-Net (Ronneberger et al., 2015). They modify the vanilla U-Net for better performance on calving front detection (Loebel et al., 2022; Mohajerani et al., 2019; Zhang et al., 2019). One approach is to segment the images into different areas and extract the calving A2 (1.8) front as the border between segmentation areas (Hartmann et al., 2021; Zhang et al., 2019; Baumhoer et al., 2019; Periyasamy et al., 2022; Loebel et al., 2022). Another approach directly trains a model on the position of the calving front (Davari et al., 2022). This task suffers from severe class imbalance due to the thin calving front. They approach this problem by creating a distance map from every pixel to the front line. The network is trained on the distance map instead of the thin front line. The actual front-line prediction is then extracted during post-processing.

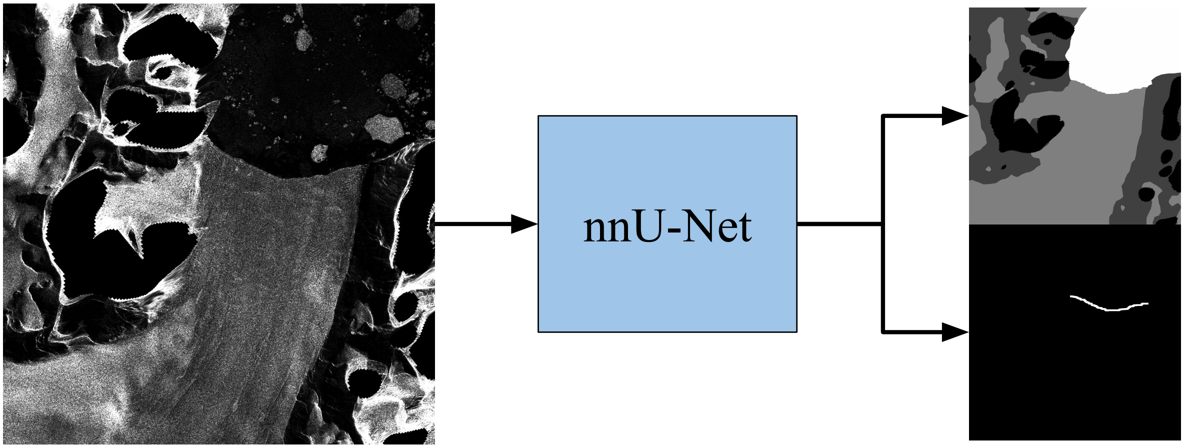


**Figure 1.** Illustration of the modified nnU-Net for simultaneous predicting landscape zones and the glacier front. On the left is an exemplary satellite image of the Crane Glacier taken by TanDEM-X (TDX) on 24 June 2011. On the right, the upper image shows the four segmentation classes: ocean (white), glacier (light grey), rock (dark grey), and no information available (black). The lower image shows the glacier's calving front as a white line versus the black background.

Other works use the segmentation network DeepLabv3 (Chen et al., 2018) to detect calving fronts. The main advantage of DeepLabv3 over U-Net is the atrous spatial pyramid pooling, which makes the network adaptable to different image resolutions (Zhang et al., 2021; Cheng et al., 2021).

Calving Front Machine (CALFIN) proposed by Cheng et al. (2021) segments optical and SAR images into the ocean-land zones and extract the calving front during post-processing with a topography map of the area. They apply MTL with a late-branching architecture. They use two labels: a binary ocean mask and a binary calving front mask. They achieve state-of-the-art predictions with an 86 m deviation from the measured calving front. A detailed comparison to aforementioned U-Net-based methods by Mohajerani et al. (2019) and Baumhoer et al. (2019) revealed the generalization ability of CALFIN on other glacier datasets. The images had to be down-sampled for CALFIN. Therefore, the distance in meters doubles but comparing the pixel distance errors reveals a similar performance. With 29 M parameters, CALFIN is still a large network that needs a large amount of training data. Chen et al. (2019) propose a similar approach for medical image segmentation and show that the individual tasks benefit from MTL. Several other approaches have advanced the U-Net architecture for MTL for medical applications (Abolvardi et al., 2020; Kholiavchenko et al., 2020; Li et al., 2019b; Amyar et al., 2020). The work of Heidler et al. (2021) uses MTL for the segmentation of a binary ocean-land mask and edge detection of the Antarctic coastline, where the calving front is just part of the coastline. They add task-specific heads for the two tasks to the U-Net. They achieve results with a deviation from the reference of 345 m compared to 483 m with the vanilla U-Net. ~~Their post-processing involves the terrain elevation, and~~To avoid distortions of the metric from areas far from the coast, the metric is calculated within 2 km of the true coastline. C5 (1.9) A further development of the HED-UNet for Antarctic ice shelf front detection is proposed by Baumhoer et al. (2023). Their A3 (1.9) post-processing includes an elevation threshold of 110 m to remove erroneous classifications in the high-altitude dry snow zones.

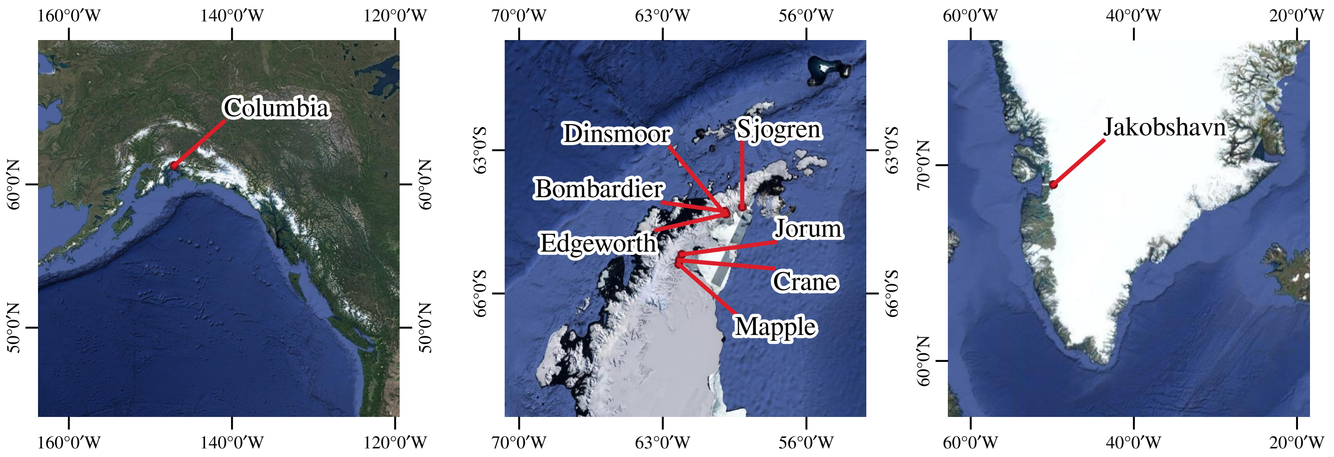


**Figure 2.** Location of the seven benchmark glaciers: Columbia Glacier in Alaska; Crane; Mapple; Jorum; Dinsmoore, Bombardier, Edgeworth (DBE); Sjogren Glacier in the Antarctic Peninsula; and Jakobshavn Isbrae Glacier in Greenland. The background images are taken from the Google Maps Satellite imagery layer © Google 2019.

| | Alaska | Antarctic Peninsula | | | | | Greenland |
|---|---|---|---|---|---|---|---|
| | Columbia | Mapple | Crane | Jorum | DBE | Sjögren-Inlet | Jakobshavn Isbrae |
| Split | - test: 122 - | | - train: 559 - | | | | |
| Images # | 65 | 57 | 69 | 77 | 133 | 121 | 159 |
| Area [km] | 32 x 15 | 8 x 8 | 19 x 25 | 20 x 13 | 22 x 20 | 23 x 19 | 16 x 19 |

**Table 1.** Properties of the dataset, including a list of captured glaciers, train-test split, number of images per glacier, and covered area.

## 3 Dataset

The dataset used in this work is provided by Gourmelon et al. (2022). It contains 681 Synthetic Aperture Radar (SAR) images of seven marine-terminating glaciers taken by six different satellites. Two glaciers are located in the northern hemisphere, namely ~~Columbia Glacier in Alaska and Jakobshavn Isbrae (Sermeq Kujalleq) in Greenland. The five glaciers in the southern hemisphere are all located on the Antarctic Peninsula (AP) (see Fig. 2). The Crane, Mapple, and Jorum glaciers are closest to the south pole, followed by Dinsmoore, Bombardier, Edgeworth. They are located so close together that they are treated as one glacier site. The last glacier is the Sjögren-Inlet Glacier.~~Columbia (COL) Glacier in Alaska and Jakobshavn Isbrae (JAK) (Sermeq Kujalleq) in Greenland. The five glaciers in the southern hemisphere are all located on the Antarctic Peninsula (AP) (see Fig. 2). The Crane, Mapple, and Jorum glaciers are closest to the south pole, followed by Dinsmoore, Bombardier, Edgeworth (DBE). They are located so close together that they are treated as one glacier site. The last glacier is the Sjögren-Inlet (SI) Glacier. C6 (2.9)

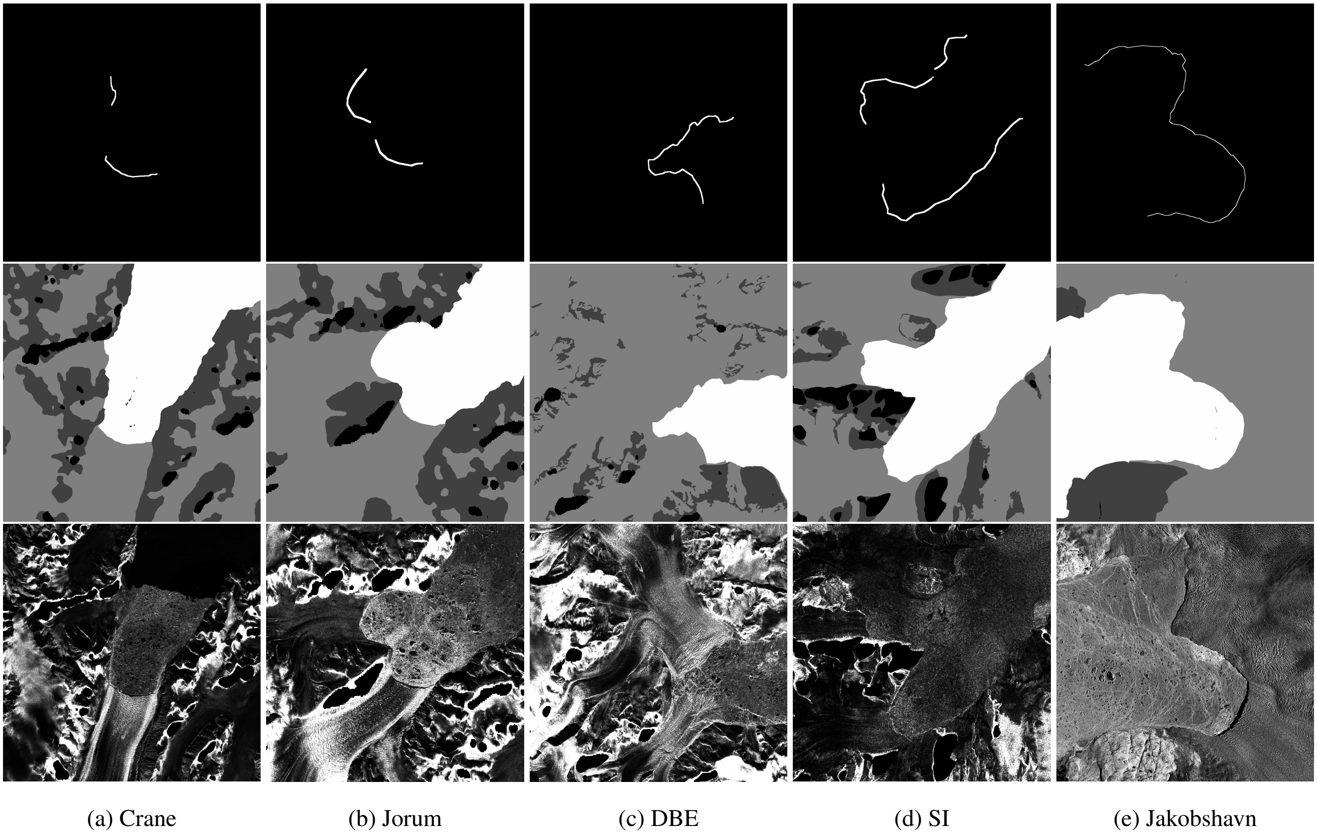


**Figure 3.** Sample images of every glacier in the training set and their corresponding labels. The first row shows the front label with a black background and a one-pixel wide white line representing the calving front. The second row contains the zone labels with four classes: ocean (white), glacier (light grey), rock (dark grey), and no information available (black). Black dots in the ocean zone of Crane are areas with no data available due to a small artefact at the former calving front of Crane Glacier in the DEM (Cook et al., 2012) used for the orthorectification of the SAR data. SAR imagery is provided by DLR, ESA, and ASF.

Table 1 lists the seven glacier sites with the number of images and the area they show. The table also depicts the train-test split. The samples of the train set are used for the parameter optimization of the segmentation method, and the test set is exclusively used for evaluating the method. The test set contains the glaciers Mapple and Columbia. Columbia is the only mountain glacier in the dataset. The other glaciers are outlet glaciers of polar ice sheets. Therefore, Columbia can be seen as a benchmark glacier for the generalization capability of the segmentation method. Columbia and Mapple also differ in the shape of the glacier. The images of Columbia show multiple calving fronts in one image, and the flow of the glacier arms goes in different directions. On the other hand, Mapple is a less complex glacier, moving in only one direction with one clearly defined calving front between the lateral fjord side walls.

The dataset contains two labels and a bounding box of the glacier for each SAR image. One shows a mask of the different zones of the glacier (ocean, glacier, no information available, rock). The other label contains a one-pixel wide line representing the calving front. The bounding box is not utilized for our method. A sample of each glacier in the training set with its corresponding labels is shown in Fig. 3. Predicting the zone mask can be seen as a classic segmentation problem. The calving front can also be extracted from the zone label by taking the border between ocean and ice areas in the corresponding bounding box. The direct delineation of the calving front is a more difficult task due to the high-class imbalance. Fewer than 1 % of the pixels are labelled as front pixels. Additionally, the structure of the class region is not a convex hull but a thin line.

Corresponding to the change of glacier area, the position of the calving front changes. In Fig. 4, the retreat of Columbia is visualized by plotting the calving front position of every sample in a different colour. The bright lines represent past calving fronts, and the dark red lines the more recent positions. The constant retreat of COL is visible through the tiered lines. The visualization of the glaciers in the southern hemisphere is included in Appendix B. The change in the class distribution of the zone labels over time is shown in Appendix. C.

Every glacier is captured by multiple satellites for a higher temporal resolution and extended observation periods. Meaning that recordings of one glacier are captured by different SAR systems with different image resolutions. The resolution refers to the ground range resolution. Environmental Satellite (ENVISAT), European Remote-Sensing Satellite-2 (ERS), Sentinel-1 (S1) has a resolution of 20 m, Phased Array type L-band SAR (PALSAR) has a 17 m resolution, and TDX has a 7 m per pixel resolution. Unfortunately, the dataset does not provide information about the polarization In Fig. 5, a timeline of the images of each glacier A4 (1.3) visualizes the observation time and frequency of the images. The first two rows show the glaciers of the test set.

## 4 Method

In this section, we explain our method, which includes the utilisation of the nnU-Net as a framework that simplifies the training of a U-Net. We document our six experimental setups that aim to evaluate the impact of MTL on the training of the U-Net.

### 4.1 Background

In the field of deep learning, a lot of time and effort is put into the hyperparameter search. Isensee et al. (2021) proposes the nnU-Net that automates the manual tuning of hyperparameters. ~~It takes a fingerprint of the dataset and adjusts hyperparameters accordingly.~~The nnU-Net is a framework around the U-Net architecture. It provides good default values for hyperparameters, rules to adapt hyperparameters to the dataset, and many established deep-learning techniques for training and inference. ~~The nnU-Net is designed with a focus on three-dimensional (3D) data like Computer Tomography (CT) and~~ C7 (2.2) ~~Magnetic Resonance Imaging (MRI), which have different spacing in dimensions. One parameter of the data fingerprint is the distribution of spacing. The annotation, image resampling strategy, and target spacing depend on the spacing distribution as rule-based parameters. Two other~~Most of the rule-based hyperparameters are only relevant for 3D data in the medical domain, like CT and MRI and are irrelevant to our experiments. Two dataset parameters that are important are intensity distribution and C8 (2.10) image modality, which determine intensity normalization. ~~If the imaging modality is CT, a global dataset percentile clipping and~~

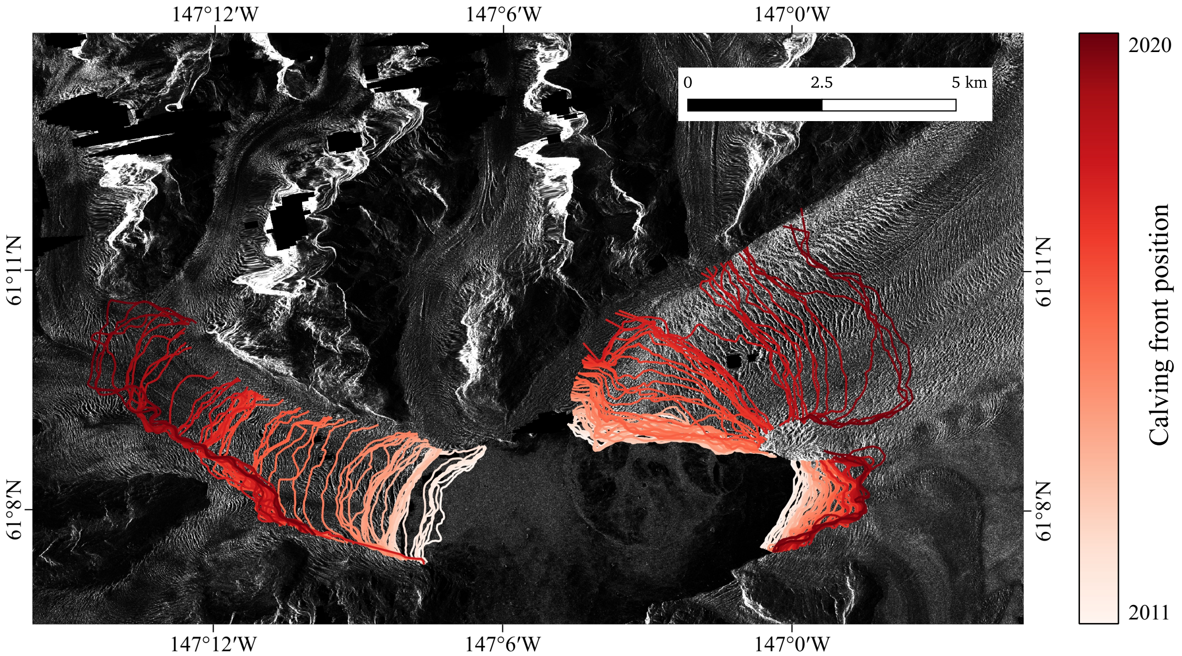


**Figure 4.** Evolution of the calving front position of Columbia Glacier (Alaska). The red lines are all calving front labels available in the dataset. Background: SAR intensity image acquired by TDX on date 13 November 2011. Projection CRS EPSG:4326 - WGS 84 / UTM zone 6N. SAR imagery is provided by DLR, ESA, and ASF.

~~z-score normalization are applied. For most samples, a large portion of the image is unlabeled, which is called the background. All the labelled areas are called foreground. For the normalization, the background is omitted, and the mean and standard deviation is calculated for the foreground. In all other cases, the~~For all image modalities except CT, z-score normalization is applied. ~~with per image mean and standard deviation.~~Each image is normalized independently by first subtracting its mean and then dividing by its standard deviation. Another dataset parameter is the median shape of the samples. The shape and the available Graphics Processing Unit (GPU) memory determine the patch and batch size and, therefore, the network topology. The patch size is initialized with the median shape and iteratively reduced while adapting the network topology accordingly until the network can be trained with a batch size of at least two given GPU memory constraints. C9 (2.10) C10 (2.11)

In addition to the rule-based parameters, there are fixed parameters. These parameters are based on the authors' experience and generalize well across various tasks. The nnU-Net uses a poly learning rate scheduler, a combination of dice coefficient and cross-entropy as the loss function, and Stochastic gradient descent (SGD) with a Nesterov momentum as optimizer. The dice coefficient also helps with the problem of class imbalance. The coefficient of each class is weighted by the number of pixels in the label that relate to the class. Additionally, deepsupervision is used to avoid vanishing gradients in neural networks with many layers. Independent of the dataset size, one epoch is defined as 250 mini-batches with foreground oversampling. The foreground oversampling is especially helpful for our application because the class imbalance between the calving front and background is A5 (2.8)

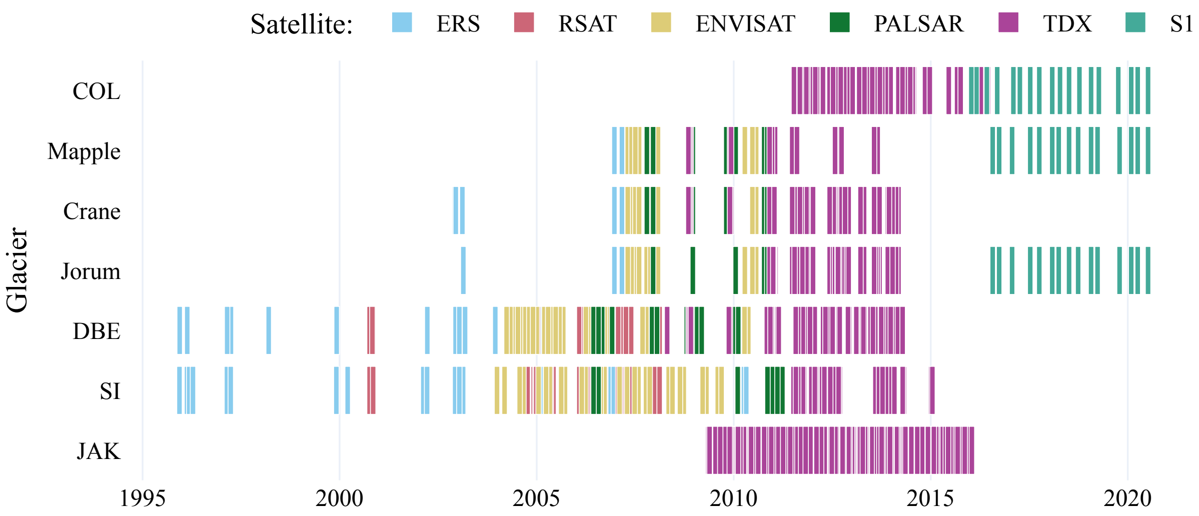


**Figure 5.** Distribution of the dataset's images over time. The samples are grouped by the seven glaciers and coloured according to the capturing satellite. We abbreviated the following glaciers: Columbia (COL), Dinsmoore, Bombardier, Edgeworth (DBE), Sjögren-Inlet (SI), Jakobshavn Isbrae (JAK).

high. It ensures that (at least) one third of the patches for training are guaranteed to contain a foreground class. In our case, every batch contains two patches, from which one is forced to contain calving front pixels. A6 (2.8)

This framework achieves robust results for various medical image segmentation tasks. Overall, nnU-Net sets a new state of the art in 33 of 53 segmentation tasks of the Kidney and Kidney Tumor Segmentation Challenge challenge (Heller et al., 2021) and otherwise shows performances on par with or close to the top leaderboard entries. But the performance of the nnU-Net on segmenting glacier SAR images is yet to be tested.

### 4.2 Multi-Task Learning with nnU-Net

The first research goal is to apply the nnU-Net out-of-the-box on the glacier front detection and on the glacier zone segmentation, respectively. Training the nnU-Net directly on the front labels is the most straightforward approach for calving front detection. The nnU-Net is intended to be used in the STL manner. In Fig. 6, these two baseline experiments are represented by the two blue columns on the left. The label of the calving front is dilated to the width of five pixels. Our preliminary experiments have shown that dilation makes the predictions more robust. For the training with zone labels, the post-processing includes extracting the boundary between the ocean and the glacier.

In the following experiments, the segmentation problem changes to a multi-task problem, where both labels are used to train one model. The next two experiments concern network architecture. They are represented by the two green columns in Fig. 6. The early-branching architecture uses one decoder for every label. Thus, the number of parameters increases by about $50\,\%$. In contrast, the late-branching architecture requires only a small change to the vanilla U-Net. An additional channel of the last

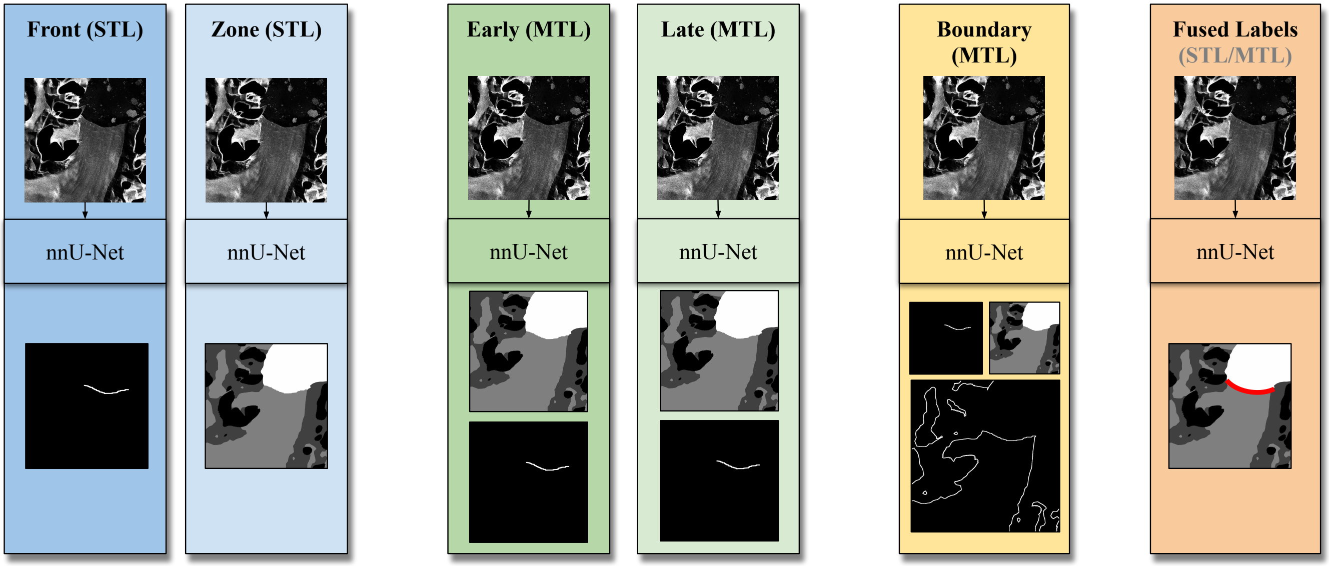


**Figure 6.** Illustration of the six experiments: Two Single-Task-Learning (STL) experiments (blue), two MTL experiments with different architectures (green), one MTL experiment with an additional task of delineating the whole glacier boundary (yellow), and one experiment where the two labels are fused into one segmentation task with an additional class in the zone label (orange).

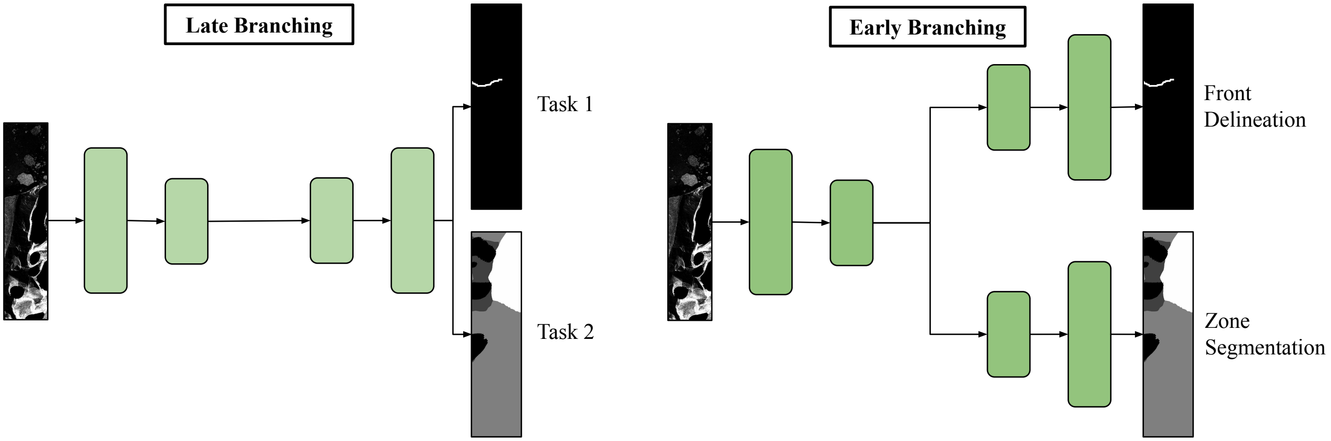


**Figure 7.** Illustration of early- and late-branching U-Net architectures for MTL. Both architectures perform joint feature extraction. The early-branching architecture applies separate reconstruction with two decoders, one for each task. The late-branching network applies joint reconstruction with a common decoder.

layer is used to predict the second label. For this architecture, only the weights of one kernel are trained additionally to the set of parameters needed for one task. The change in the total number of parameters that need to be trained is negligible.

Because architecture changes with multi-task learning were not foreseen in nnU-Net framework, we had to make changes to the framework. During the experiment's planning, i. e., pre-processing phase, we fixed the network architecture's estimated size so that late- and early-branching networks return the same value for the network size. Otherwise, early and late branching networks would be trained on different patch sizes. Thus, performance differences during the evaluation might arise from the different patch sizes and not the differing architectures. During training, the error of all labels is calculated and summed up with equal weighting. ~~Only minor changes had to be made to the inference part.~~For the inference, we adapted only the number of channels in the case of MTL. C11 (2.12) After the test samples are divided into patches and fed through the network and the patch predictions are combined into the prediction of the whole image. The predictions of the zones are post-processed to get an additional result for the position of the calving front. All glacier pixels with a neighbouring pixel classified as the ocean are classified as glacier front to extract the glacier front from the zone predictions.

The last two experiments concern label changes. In Fig. 6, they are coloured yellow and orange. The fifth experiment of this work, cf. Fig. 6 (yellow), extracts the boundaries between the glacier zone and all other zones as a third segmentation task for the late-branching U-Net. The label of the glacier boundaries was extracted from the zone label. All glacier pixels with a neighbouring rock or shadow pixel are assigned as glacier boundaries. The hypothesis is that providing more information about the same sample benefits the performance of the U-Net on the individual tasks. The third segmentation task is not considered in the final evaluation. A7 (2.13) The last experiment fuses the zone and front labels by creating a fourth class in the zone label associated with the glacier front, cf. Fig. 6 (orange). As the front line has a width of five pixels ($35-100\,\mathrm{m}$ depending on the image resolution), the other zone classes are merely impaired. A post-processing step is added to the predictions. The front pixels are isolated to generate a front prediction for comparing the results with the other experiments. To generate a prediction of the zones, pixels classified as front are assigned to the ocean, and the glacier zone is dilated once with a 7x7 kernel.

### 4.3 Experimental Setup

The nnU-Net was trained on a NVIDIA RTX 3080 with 12GB memory; the adapted network architecture has nine encoder blocks and eight decoder blocks. Each block consists of two convolutional layers: An instance normalization and a Rectified Linear Unit (ReLU). The kernels of all convolutional layers have a size of 3x3. During training, one batch contains two images. The patch size of the experiments that include only one label is 1280x1024. The experiments that have more than one label have a patch size of 1024x896 because of the GPU memory limit. There are also fixed parameters that are independent of the dataset. This includes the SGD optimizer with an initial learning rate of 0.01, a Nesterov momentum of 0.99, and a weight decay of 3e-5. Training of one epoch took between 100 s and 160 s. The maximum number of epochs of 500 is reached in every training (due to limited resources, we reduced the original maximum number of epochs from 1000 to 500). The nnU-Net defines one epoch using a fixed number of iterations (250). In each iteration, the batch is sampled depending on the class distribution of the sample to counteract the class imbalance. A8 (1.10, 2.2)

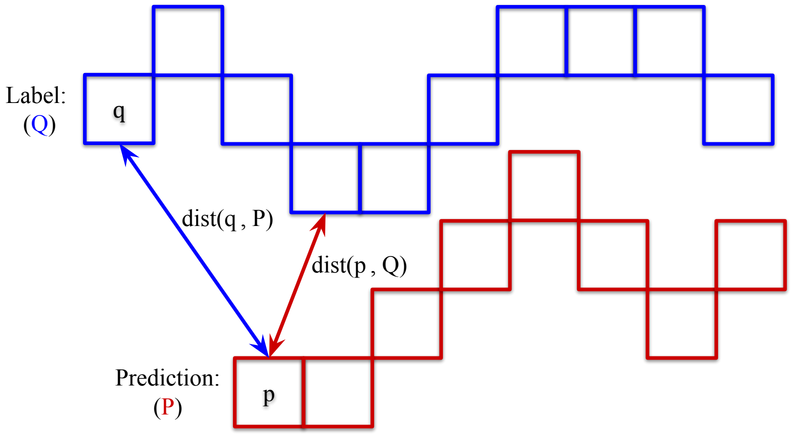


**Figure 8.** Visualization of the Mean Distance Error (MDE) calculation between front label $Q$ and front prediction $P$. The $dist(q,P)$ represents the distance between $q$ and the pixel in $P$ that is closest to $q$.

## 5 Evaluation

In this section, we will examine our evaluation metrics, compare the results of the six proposed experiments and evaluate the results of the fused label experiment. We used a five-fold cross-validation for the evaluation to eliminate the weight initialisation bias and the data-split bias into training and validation sets. The metric scores of the individual models are averaged to get a robust measure independent of weight initialization and split.

### 5.1 Evaluation Metrics

As our main measure, we use the Mean Distance Error (MDE) proposed by Gourmelon et al. (2022). It measures the distance of the predicted front $\mathcal{P}$ to the ground truth calving front $\mathcal{Q}$ for all images in the test set $\mathcal{I}$. For every pixel in the label front $\mathcal{Q}$, the distance to the closest pixel in the predicted front $\mathcal{P}$ is determined. Additionally, to make the metric symmetric, the distance to the closest pixel in the label front $\mathcal{Q}$ is determined for every pixel in the predicted front $\mathcal{P}$. These distances are averaged and taken as the mean distance between the two lines (see Fig. 8 and Eq. 2). We note that the front MDE is also calculated for the zone segmentation. We extract the front from the zone segmentation by assigning all glacier pixels with neighbouring ocean pixels as the glacier front. A9 (2.14)

$$\mathrm{MDE}(\mathcal{I}) = \frac{1}{\sum_{(\mathcal{P},\mathcal{Q})\in\mathcal{I}}(|\mathcal{P}|+|\mathcal{Q}|)} \sum_{(\mathcal{P},\mathcal{Q})\in\mathcal{I}} \left( \sum_{p\in\mathcal{P}} \min_{q\in\mathcal{Q}} ||p-q||_2 + \sum_{q\in\mathcal{Q}} \min_{p\in\mathcal{P}} ||p-q||_2 \right) \tag{2}$$

Additionally, we give classical segmentation metrics to evaluate the zone prediction. They include the Intersection over Union (IoU), which is defined as $\mathrm{IoU} = \frac{T_P}{T_P+F_P+F_N}$, i. e., the True Positive ($T_p$) pixels over the sum of $T_p$, False Positive ($F_p$), and False Negative ($F_N$) pixels. Additionally, the F1-score is computed ($F1 = 2 \cdot \frac{pr \cdot re}{pr+re}$), which is a combination of Recall

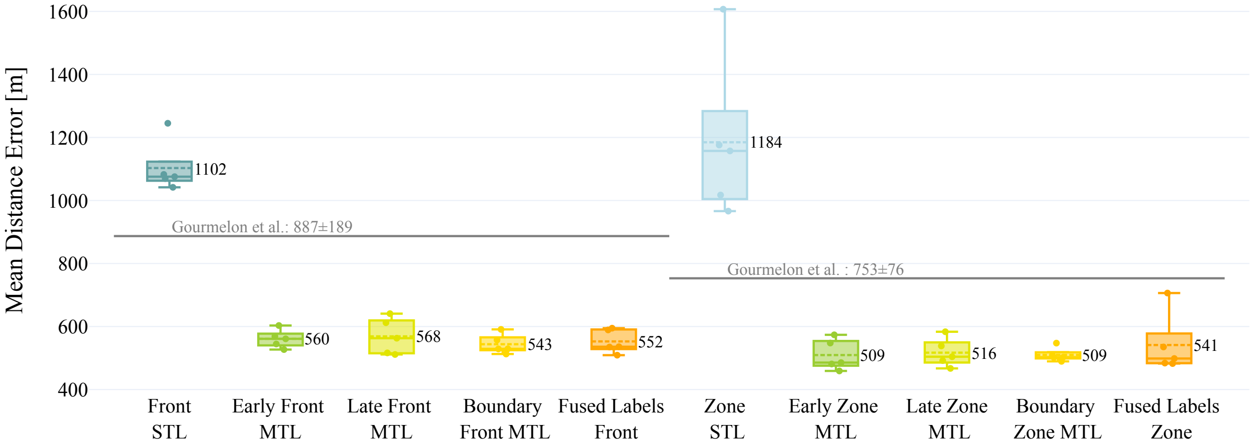


**Figure 9.** Box plot of the MDE of the six experiments. Two columns are coloured equally for each experiment except for the STL experiments. The left half represents the front delineation based on the front segmentation (Task 1), and the right half represents the front delineation based on the zone segmentation (Task 2). The baseline of Gourmelon et al. (2022) is displayed as a grey line in each half.

($re = \frac{T_p}{T_p+F_N}$) and Precision ($pr = \frac{T_P}{T_P+F_P}$). Because this segmentation task is not binary, but every sample has four classes, the metrics are calculated for each class individually and then averaged with equal weighting.

## 5.2 Results and Discussion

In this section, we evaluate our experiments. In section 5.2.1, we compare the performance of the six experiments, which we described in section 4.2. In section 5.2.2, we analyse the prediction of the fused label experiment and investigate the influence of season, glacier site and satellite on the calving front prediction performance.

### 5.2.1 Comparison of the five experiments

A10 (1.1)

In Fig. 9, the MDE of every experiment is compared. The STL approach that is trained on the front labels has an MDE of $1103 \pm 72\,\mathrm{m}$ and the STL approach that is trained on the zone labels has an MDE of $1184 \pm 225\,\mathrm{m}$. A difference between the STL experiments arises in the performance variance, where the model trained on the zone labels is larger. The number of samples that are incorrectly predicted to have no front pixel is similar, with an average of $27.2 \pm 6.0$ for training on the front and $24.8 \pm 12.4$ out of 122 for training with zone labels.

The lowest number of samples with false non-front-detections achieves the model that is trained on the fused labels, $3.2 \pm 2.0$ for the front label out of 122 samples. The baseline of Gourmelon et al. is $1 \pm 1$ based on the zone segmentation and $7 \pm 3$ from the front prediction. All values can be seen in Table A2. All MTL models have a significantly lower MDE than STL models with a significance level of $\alpha = 0.01$ using a student's t-test. The Table with the T-values of all experiment pairs is given in Table A1. The model that is additionally trained on the glacier boundary has the smallest MDE for both tasks. The MDE baseline of Gourmelon et al. is $887 \pm 189\,\mathrm{m}$ for the front prediction and $753 \pm 76\,\mathrm{m}$ for the zone. Overall the MDE is similar for all MTL approaches. The student's t-test shows that the differences between fused label and all other MTL approaches are insignificant ($\alpha = 0.33$).

The metrics for the zone segmentation, shown in Table A3, show a similar trend. An improvement of all MTL over the STL approach and minor changes between the MTL approaches. The STL of the nnU-Net on zone label achieves $62.4 \pm 3.5$ IoU and $71.7 \pm 3.2$ F1-score. The highest F1-score achieves the model that is trained with the additional boundary task with $81.7 \pm 0.5$. The model's IoU is $72.6 \pm 0.4$. The baseline of Gourmelon et al. is $80.1 \pm 0.5$ F1-score and $69.7 \pm 0.6$ IoU.

The fused label approach is the most feasible method discussed as follows. ~~This method performs similarly well in comparison to other MTL approaches while using a minimum number of parameters and using the nnU-Net framework with no profound changes. Moreover, this approach has a relatively small number ($5 \in 122$) of predictions with falsely non-detected fronts.~~ This approach performs on par with the other experiments. Even though the MDE of the boundary experiment is slighty lower than for the fused label approach, the student t-test showed that the difference is not significant. Moreover, the fused label approach requires less parameters and no changes to the nnU-Net framework, making it's use truly out-of-the-box. Additionally, this approach has a relatively small number ($5 \in 122$) of predictions with falsely non-detected fronts. C12 (1.11)

### 5.2.2 Analysis of the fused label experiment

A11 (1.1)

For the final evaluation, the zone segmentations of the five models that are trained during the five-fold cross-validation on A12 (1.13) the fused labels are combined into a more robust ensemble prediction. The models differ by their weight initialization and train-validation split. But are all trained on the fused labels. The ensemble prediction is created by taking the mean of the A13 (1.13) probabilities for every class each model gives. We don't use the calving front pixels directly from the prediction but apply the post-processing of the zone label to gain a slightly better MDE. The post-processing labels the edge of the glacier zone as a calving front that has neighbouring ocean pixels. The performance of the ensemble prediction results in an MDE of $515 \pm 39\,\mathrm{m}$ A14 (2.14) and $5 \in 122$ predictions with no fronts, which is similar to the average performance of the single models. We did not use the ensemble prediction for the comparison between the experiments in section 5.2.1, to show the variance that is introduced by a different train-validation split and different weight initialization. A15 (1.13)

The distribution of the MDE on the test set predictions is plotted in Fig. 10. In the first row, all errors of the prediction of the 117 test samples are drawn as dots. The test set contains two glaciers: Mapple and COL. The rows below show the distribution of the MDE on the two test glaciers separated into summer and winter seasons. The main difference in MDE is between glaciers, similar to the baseline results of Gourmelon et al. (2022) with MDE of $287 \pm 48\,\mathrm{m}$ for Mapple and $840 \pm 48\,\mathrm{m}$ for COL. The MDE of our methods is on average $109 \pm 90\,\mathrm{m}$ for Mapple, while the MDE of COL is $930 \pm 1420\,\mathrm{m}$. The difference of the MDE is caused by a group of predictions with an error $> 1000\,\mathrm{m}$. The median value is $222\,\mathrm{m}$ for COL and $80\,\mathrm{m}$ for Mapple. The reasons for the large glacier differences might be related to the different shapes. Mapple has a simple calving front, a single

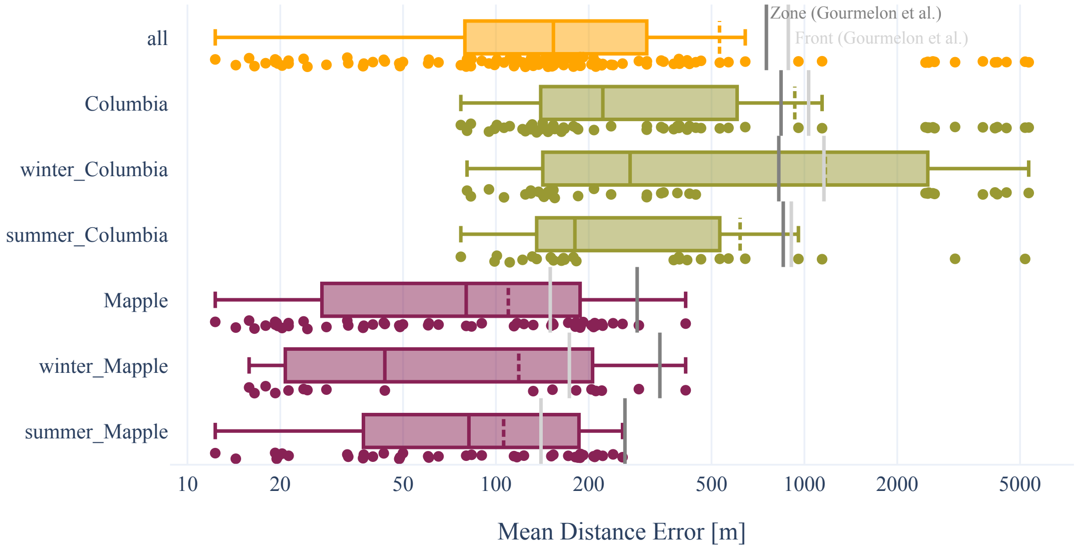


**Figure 10.** MDE on the test set grouped by glaciers and subdivided into seasons. The orange box plot in the first row shows all MDE of the test set. The olive green plot below shows the errors from the COL images, subdivided into seasons in the third and fourth rows. The last three rows show errors from the Mapple images. The dark grey line shows the baseline of the zone prediction, and the light grey line shows the baseline of the front prediction from Gourmelon et al. (2022). The ensemble model was trained with the fused zone and front labels. The y-axis has a logarithmic scale. The median is the middle line in the rectangle, and the dashed line represents the mean. The x-axis has a logarithmic scale. Otherwise, the outliers would dominate the plot. The rectangle reaches from the first quartile to the third quartile.

line constraint by a straight valley. Columbia has multiple calving fronts, a stream coming from the left side of the image, one from the top, and another from the left (see Fig. 13).

There is also a seasonal difference in the MDE. The MDE of the front prediction during the summer of Mapple and COL combined have lower MDE $307 \pm 730\,\mathrm{m}$ than the samples captured during the winter months ($818 \pm 1389\,\mathrm{m}$). However, the medians are closer together with $150\,\mathrm{m}$ in the summer months and $181\,\mathrm{m}$ in the winter months. The MDE baseline (Gourmelon et al., 2022) in summer is $732 \pm 93\,\mathrm{m}$ and $776 \pm 65\,\mathrm{m}$ in winter. Although, most outliers are from winter seasons in COL. In winter, snow coverage of the glacier is more likely ~~, covering~~. The SAR signal can penetrate through several meters of snow cover, depending on the SAR frequency and the water content of the snow cover. However, most of the studied glaciers, in particular Columbia Glacier and the glaciers on the Antarctic Peninsula, are located in temperate marine environments making wet snow conditions also likely during winters. Thus snow can cover useful artefacts like crevasses and rock structures that can C13 (1.12, 1.1) be useful for the pattern recognition of the nnU-Net. ~~Additionally, the ocean is covered more often with ice mélange, which reduces the contrast between ocean and glacier areas.~~However, more important is the fact that the ocean next to the glaciers is covered more often by ice mélange during winters, which reduces the contrast between ocean and glacier areas. Even for experienced human mappers, it can be a challenging task to distinguish between the glacier tongue and the ice mélange. Thus, we hypothesize that the network has similar issues leading to reduced performance during winter. The difference between C14 (1.12, 1.1) seasonal MDE also depends on the sensor. The MDE of the low-resolution sensors (S1, ENVISAT) is lower in the summer month. The MDE of high-resolution sensor (TDX) is lower in winter months but the difference is much smaller. A table of the MDE grouped by satellite and season is in Appendix A4. A16 (1.2)

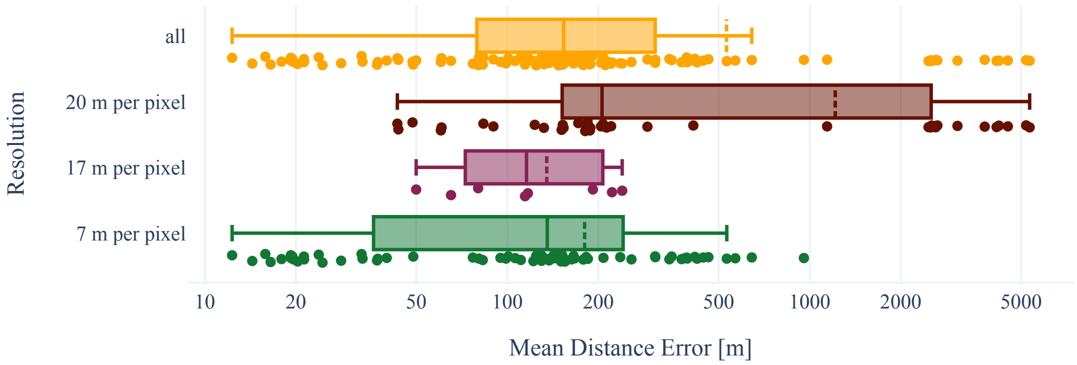


**Figure 11.** MDE on the test set grouped by resolution of the SAR image. The orange box plot in the first row shows all MDE of the test set. ENVISAT, ERS and S1 have a resolution of 20 m per pixel. MDE of this resolution are represented by the dark red box plot. PALSAR has a 17 m per pixel resolution. MDE of images of this resolution are represented in the magenta box plot. TDX has a 7 m per pixel resolution. MDE of images of this resolution are represented in the green box plot. The ensemble model was trained with the fused zone and front labels. The dark grey line shows the baseline of the zone prediction, and the light grey line shows the baseline of the front prediction from (Gourmelon et al., 2022). The y-axis has a logarithmic scale. The median is the middle line in the rectangle, and the dashed line represents the mean. The x-axis has a logarithmic scale. Otherwise, the outliers would dominate the plot. The rectangle reaches from the first quartile to the third quartile.

An overview of the impact of the sensor resolution is given in Fig. 11. The images with a resolution of 7 m per pixel have a mean MDE of 180 m resolution have a mean MDE of 135 m, but this class contains only images of Mapple glacier taken by PALSAR. The images with 20 m per pixel have a mean MDE of 1214 m. The distribution has a cluster of MDEs that is larger than 1000 m. These large errors only stem from images of COL (see Fig. 12 row Columbia_S1). A17 (1.3)

In Fig. 12, the MDE is grouped by satellites. The predictions for the samples captured by ERS, ENVISAT, ~~PALSAR, and TDX~~and PALSAR have a similar average error between 150 m and 300 m. C15 (1.14) However, the samples from ERS, ENVISAT, and PALSAR only represent Mapple Glacier. TDX captures both test sites an has a MDE of 180 ± 179 m The samples captured by A18 (1.14) S1 have a higher MDE with 1664 ± 1807 m. The outliers are all from samples captured by S1 of Columbia. The neural network can not generalise from the training set to this set of samples. The 14 outliers have a front-line delineation error of $> 1000$ m. These are predictions from images of COL taken by S1. The low-resolution, e. g., cracks in the ice, are not visible. Fig. 13 (c) shows that the model falsely predicts ocean pixels as the front and does not predict the left calving front at all. The MDE for this sample is 4221 ± 5026 m. We suspect that the complex calving front of COL requires a high resolution for accurate detection A19 (1.17) or more training data of S1. The group of outliers heavily increases the overall MDE but can be separated clearly by specifying A20 (1.15)

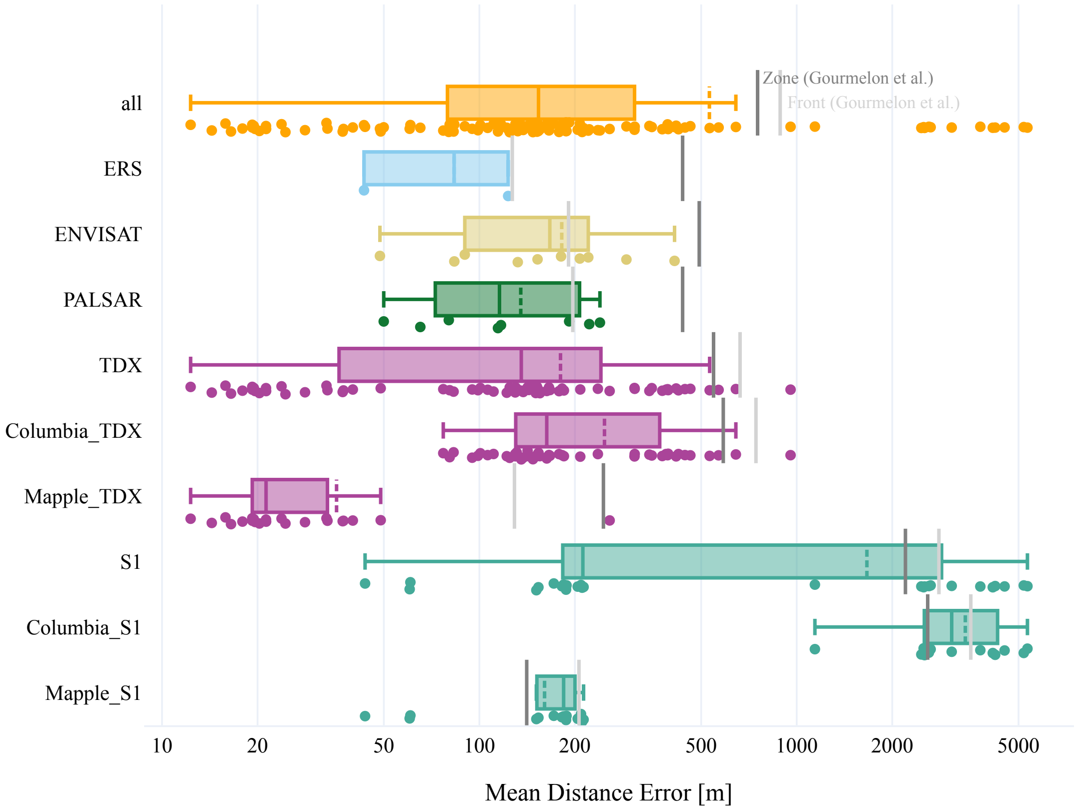


**Figure 12.** MDE on the test set grouped by satellites. The dark grey line shows the baseline of the zone prediction, and the light grey line shows the baseline of the front prediction from (Gourmelon et al., 2022). The images from ERS (light blue), ENVISAT (yellow), and PALSAR (green) only capture the Mapple Glacier. The MDE from TDX (pink) and S1 (turquoise) is subdivided into the two glaciers of the test set. The ensemble model was trained with the fusion of zone and front labels. The y-axis has a logarithmic scale. The median is the middle line in the rectangle, and the dashed line represents the mean. The x-axis has a logarithmic scale. Otherwise, the outliers would dominate the plot. The rectangle reaches from the first quartile to the third quartile.

satellite and glacier. It shows the generalisation limits of our method. For a visual inspection, we provide the predictions of the test set on Zenodo (https://zenodo.org/record/7915740#.ZFtS3dIzZhE). A21 (1.15)

On the other hand, the model predictions of the calving front have a high overlap with the label on images of Mapple with a 330 resolution of ≥ 20 m (ERS) (see Fig. 14). There are some outliers for Mapple as well, but they are not as severe as the outliers at COL. The prediction of Mapple with a high resolution of 7 m per pixel is close to the ground truth (Fig. 14).

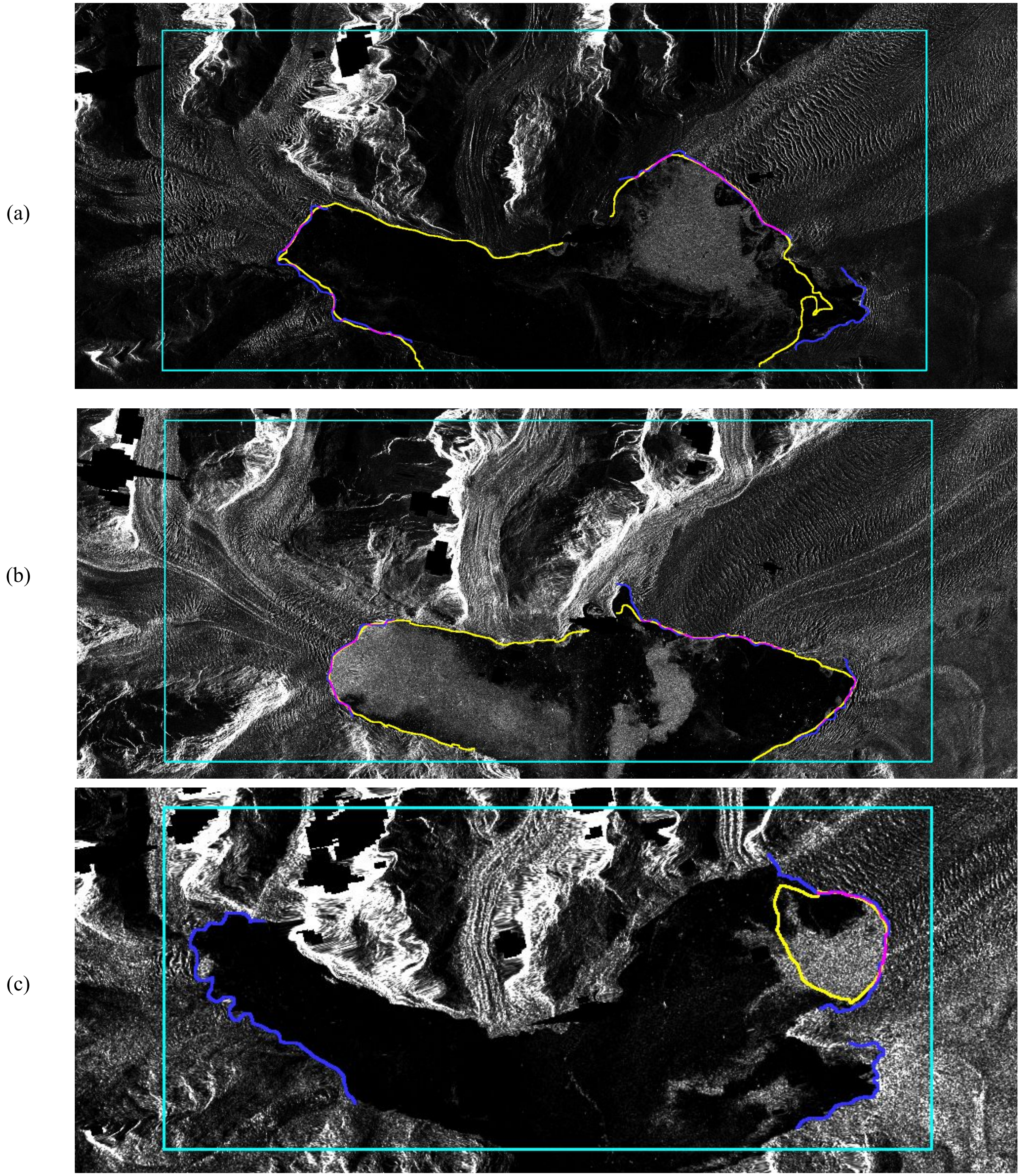
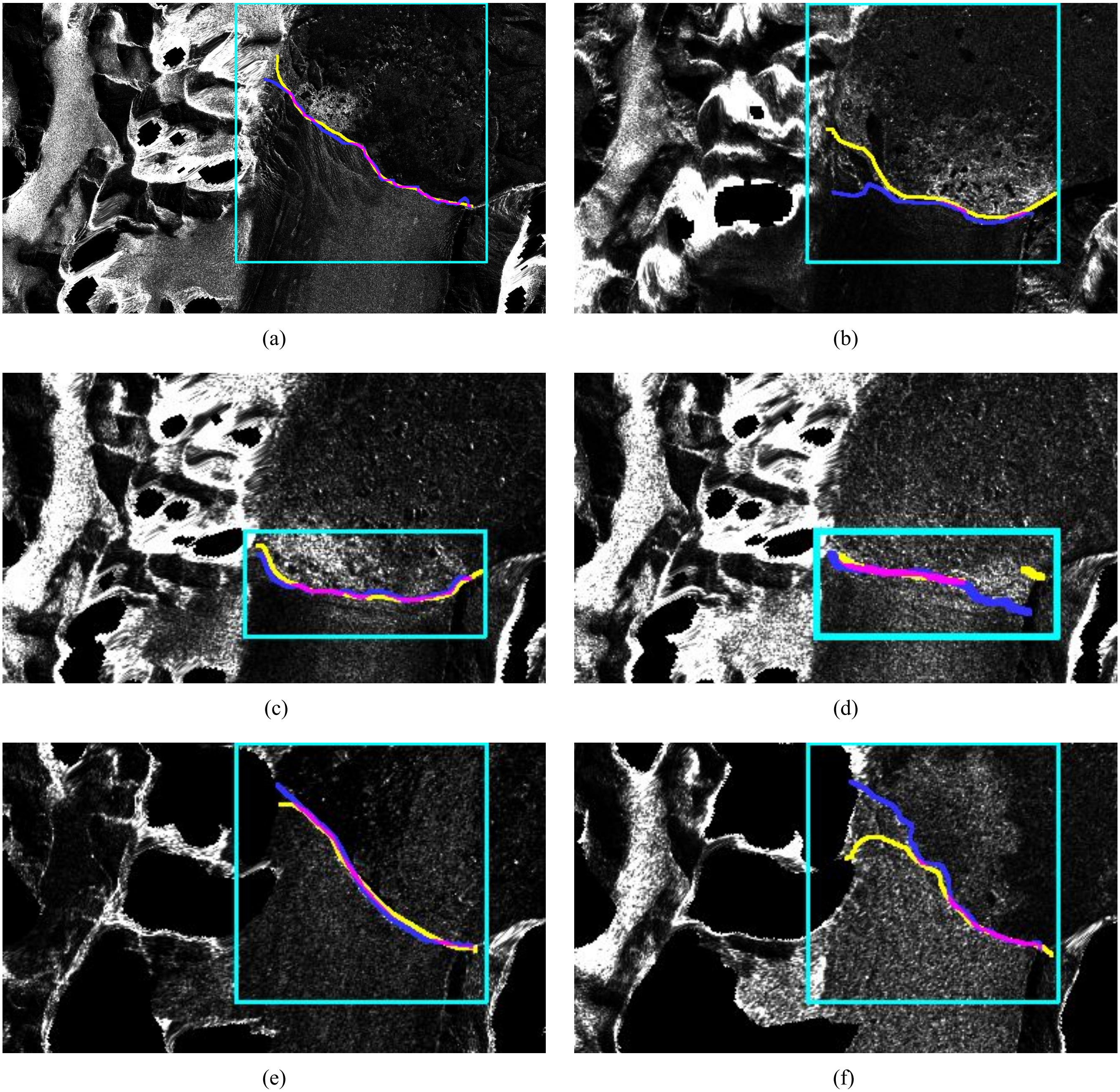


**Figure 13.** Calving front predictions of COL on (a) 22 June 2014, and (b) 11 February 2016 taken by TDX with 7m pixel resolution (after multi-looking with a factor of 3x3); (b) was taken by S1 on 10 September 2019 with 20 m pixel resolution; ground truth (blue), prediction (yellow), the overlap of ground truth and prediction (magenta). SAR imagery is provided by DLR, ESA, and ASF.



(a)

(b)

(c)

(d)

(e)

(f)

**Figure 14.** Calving front predictions of Mapple Glacier; ground truth (blue), prediction (yellow), the overlap of ground truth and prediction (magenta), bounding box (cyan). (a) was taken by TDX on 13 October 2008. (c) was taken by PALSAR on 29 November 2008. (c) was taken by S1 on 3 July 2016. (d) was taken by S1 on 2 March 2019. (e) was taken by ERS on 5 February 2007. (f) was taken by ENVISAT on 22 September 2007. SAR imagery is provided by DLR, ESA, and ASF.

## 6 Conclusion

This work explores the use of the nnU-Net by Isensee et al. (2021) to segment glacier calving fronts. The nnU-Net promises an out-of-the-box application of deep learning for segmentation tasks. We evaluate this claim on a dataset of glacier images provided by Gourmelon et al. (2022). The dataset contains two tasks. The calving front detection and the glacier zone segmentation. We try different modifications of Multi-Task-Learning with two different neural network architectures to tackle both tasks simultaneously. The results show that combining both tasks ~~benefits~~increases each task's performance. C16 (1.18) ~~Still, there is no significant difference between the two network architectures~~No significant difference between the two Multi-Task-Learning architectures exists. C17 (1.4) Adding more domain-specific tasks like glacier boundary delineation does not further improve the previous tasks. Due to the small area of the front line, the two labels can be fused into one label. The fusion of labels decreases the number of parameters used, shortens the training time, and reduces the deep learning expertise needed, as the nnU-Net can be used without modifications. With this approach, we achieve an average MDE of $541 \pm 84\,\mathrm{m}$. We provide the code and the pre-trained model for application on further SAR images of glacier fronts (see Code and data availability). The predictions need to be filtered manually since there can be outliers, as our results show. But it will provide an initial prediction, which eases the task of glacier front delineation. A22 (1.4)

To improve the average MDE, future work should focus on reducing the model performance on low-resolution images like S1 with $20\,\mathrm{m}$ per pixel resolution. A23 (1.18) This can be done by including more images taken by S1 in the training set or by implementing an oversampling strategy of the low-resolution images. A24 (1.4)

The framework nnU-Net is well suited for segmenting SAR images of glaciers and calving front delineation. The modification of the nnU-Net for MTL improves the results compared to STL. However, it is not necessary for glacier segmentation and calving front delineation because both labels can be fused without losing a significant amount of information. Our findings highlight the suitability of the nnU-net for glacier front segmentation on multi-mission SAR remote sensing data, which will facilitate an efficient, extended spatiotemporal mapping of tidewater glacier terminus changes. Our findings also promote the out-of-the-box application of the nnU-Net for other segmentation tasks based on satellite imagery, because we didn't need to modify it for the calving front detection. A25 (1.4)

## Appendix A: Evaluation results of all experiments

This section provides more detailed values of the evaluation. Table A1 contains the T-values for stating significant and insignificant differences between models. Table A2 contains the zone MDE and the front MDE of the six experiments. The MDE is averaged over the five model results. Additionally, the Baseline from (Gourmelon et al., 2022) is provided in the first two rows. The values correspond to Fig 9. We provide Fig. table A5 to compare the two methods (nnU-Net and (Gourmelon et al., 2022)). Table A4 contains the MDE of the fuse label training grouped by satellite and season.

| | Early | Late | Boundary | Fused |
|---|---|---|---|---|
| | MTL | MTL | MTL | Labels |
| STL | 15.78 | 13.46 | 16.14 | 14.29 |
| Early MTL | 0 | 0.31 | 1.02 | 0.43 |
| Late MTL | | 0 | 0.95 | 0.58 |
| Bound. MTL | | | 0 | 0.46 |

(a) Front position extracted from front prediction.

| | Early | Late | Boundary | Fused |
|---|---|---|---|---|
| MTL | MTL | MTL | Labels | |
| STL | 6.57 | 6.51 | 6.66 | 5.96 |
| Early MTL | 0 | 0.29 | 0.03 | 0.75 |
| Late MTL | | 0 | 0.35 | 0.58 |
| Bound. MTL | | | 0 | 0.81 |

(b) Front position extracted from zone prediction.

**Table A1.** T-value for the significance of difference. T-values that surpass the threshold of 3.36 means that the probability for the difference to be random is below 1 %. T-values below 1 mean that the difference is by chance, with a probability of 35 %, meaning that the difference is insignificant.

| Model | Experiment | Modality | MDE Front ↓ [m] | ∅ Front | MDE Zone ↓ [m] | ∅ Zone |
|---|---|---|---|---|---|---|
| Gourmelon et al. | Front | STL | 887 ± 189 | 7 ± 3 | - | - |
| | Zone | STL | - | - | 753 ± 76 | 1 ± 1 |
| nnU-Net | Front | STL | 1102 ± 72 | 27.2 ± 6.0 | - | - |
| | Zone | STL | - | - | 1184 ± 255 | 24.8 ± 12.4 |
| | Early | MTL | 560 ± 25 | 8.6 ± 3.0 | 509 ± 43 | 2.2 ± 0.4 |
| | Late | MTL | 568 ± 51 | 8.4 ± 3.5 | 516 ± 40 | 4.0 ± 1.1 |
| | Boundary | MTL | 543 ± 27 | 5.6 ± 1.4 | 509 ± 19 | 4.4 ± 1.8 |
| | Fused | MTL | 552 ± 33 | 3.2 ± 2.0 | 541 ± 84 | 3.4 ± 1.3 |

**Table A2.** MDE of all six experiments. Every experiment setup is trained five times with different weight initialization and train-validation-split and then averaged. The  columns show the number of images for which falsely no front was detected out of the 122 test samples.

## Appendix B: Calving front labels

The other six glacier sites in this section are displayed complementary to Fig 4. The corresponding calving fronts of the different timesteps are colored with a gradient from bright for the past and dark red for the most recent fronts.

## Appendix C: Temporal distribution of zone label

Figure C1 show the temporal distribution of the zone label. JAK and COL, the two glaciers in the northern hemisphere, show an increase in ocean area with a decreasing glacier area. In particular, the set of images of the glaciers on the AP has samples with large areas with no available information. Low partial coverage by the radar swath causes prominent peaks of the no

| Model | Experiment | Modality | Precision↑ | Recall↑ | F1↑ | IoU↑ |
|---|---|---|---|---|---|---|
| Gourmelon et al. | Zone | STL | 84.2 ± 0.5 | 79.6 ± 0.9 | 80.1 ± 0.5 | 69.7 ± 0.6 |
| nnU-Net | Front | STL | - | - | - | - |
| | Zone | STL | 81.2 ± 2.4 | 71.7 ± 3.2 | 71.9 ± 3.7 | 62.4 ± 3.5 |
| | Early | MTL | 87.4 ± 0.4 | 80.9 ± 0.7 | 81.6 ± 0.6 | 72.6 ± 0.9 |
| | Late | MTL | 86.4 ± 0.4 | 79.9 ± 0.6 | 80.7 ± 0.7 | 71.1 ± 0.7 |
| | Boundary | MTL | 87.5 ± 0.3 | 80.8 ± 0.4 | 81.7 ± 0.5 | 72.6 ± 0.4 |
| | Fused | MTL | 87.0 ± 0.2 | 79.1 ± 1.7 | 80.7 ± 0.1 | 70.8 ± 1.8 |

**Table A3.** Segmentation metrics of all six experiments. Every experiment setup is trained five times with different weight initialization and train-validation-split. The metric results of each run are then averaged.

| Satellite: | S1 | ENVISAT | ERS | PALSAR | TDX | all |
|---|---|---|---|---|---|---|
| winter | 2529 ± 1719 m | 241 ± 101 m | - | - | 160 ± 125 m | 818 ± 1389 m |
| summer | 798 ± 1442 m | 122 ± 61 m | 83 ± 39 m | 135 ± 68 m | 197 ± 213 m | 307 ± 730 m |

**Table A4.** MDE of the fused label experiments grouped by sensor and season.

information available class. JAK's area distribution shows a repetitive structure of increasing and decreasing glacier area. This pattern represents seasonal changes.

| Hyperparameter | Gourmelon et al. (2022) | nnU-Net (Isensee et al., 2021) |
|---|---|---|
| Number of convolutional layers | 10 | 34 |
| Pooling | Max-pooling | Strided convolution |
| Activation-function | ReLU | Leaky ReLU |
| Patch size | 256x256 | 1280x1024 |
| Batch size | 16 | 2 |
| Optimizer | cyclic learning rate scheduler (Smith, 2017) in combination with the Adam optimizer (Kingma and Ba, 2014) | SGD with a Nesterov momentum of 0.99, and a weight decay of 3e-5 |
| Initial learning-rate | $4 \cdot 10^{-5}$ | $1 \cdot 10^{-3}$ |
| Gradient clipping | True | False |
| Deep-supervision | False | True |
| Loss-function | Combination of Dice and Cross-entropy | Combination of Dice and Cross-entropy |

**Table A5.** Comparison of the U-Net training setup and hyperparameters between Gourmelon et al. (2022) and the nnU-Net (Isensee et al., 2021).

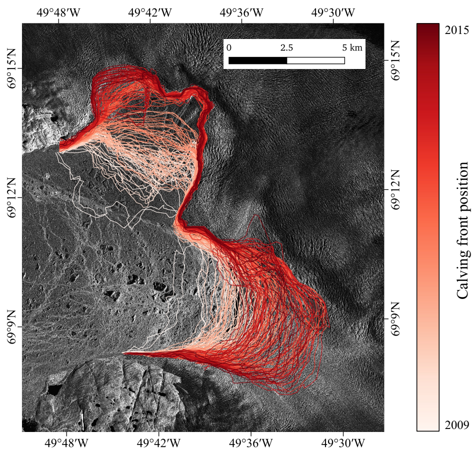


**Figure B1.** Jakobshavn Isbrae Glacier from 1995 to 2014.

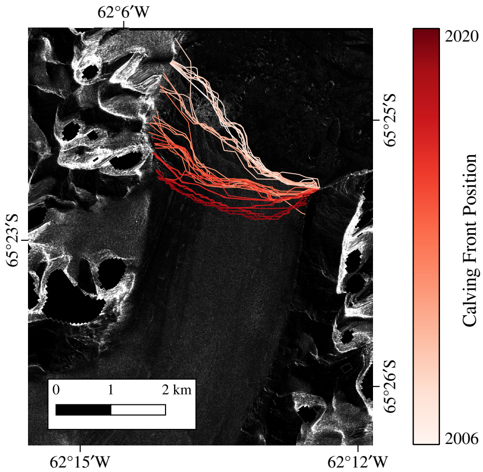


**Figure B2.** Mapple Glacier from 2006 to 2020.

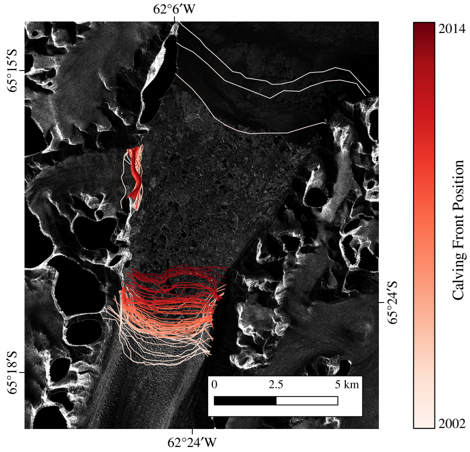


**Figure B3.** Crane Glacier from 2002 to 2014.

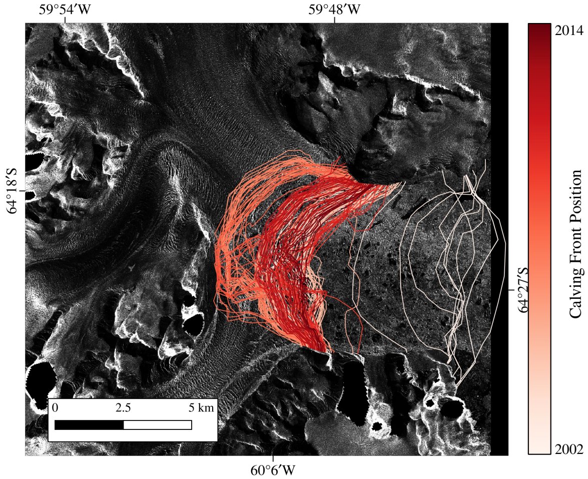


**Figure B4.** DBE Glacier from 1995 to 2014.

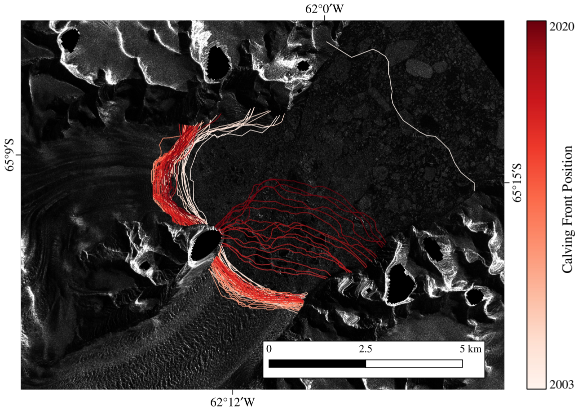


**Figure B5.** Jorum Glacier from 2003 to 2020.

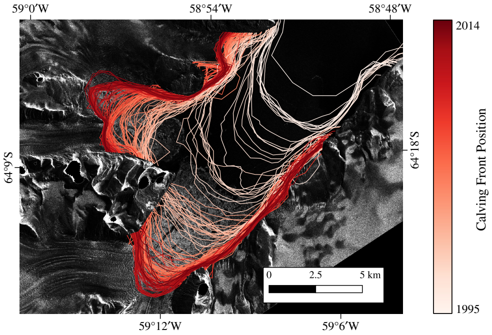


**Figure B6.** Sjögren-Inlet Glacier from 1995 to 2014.

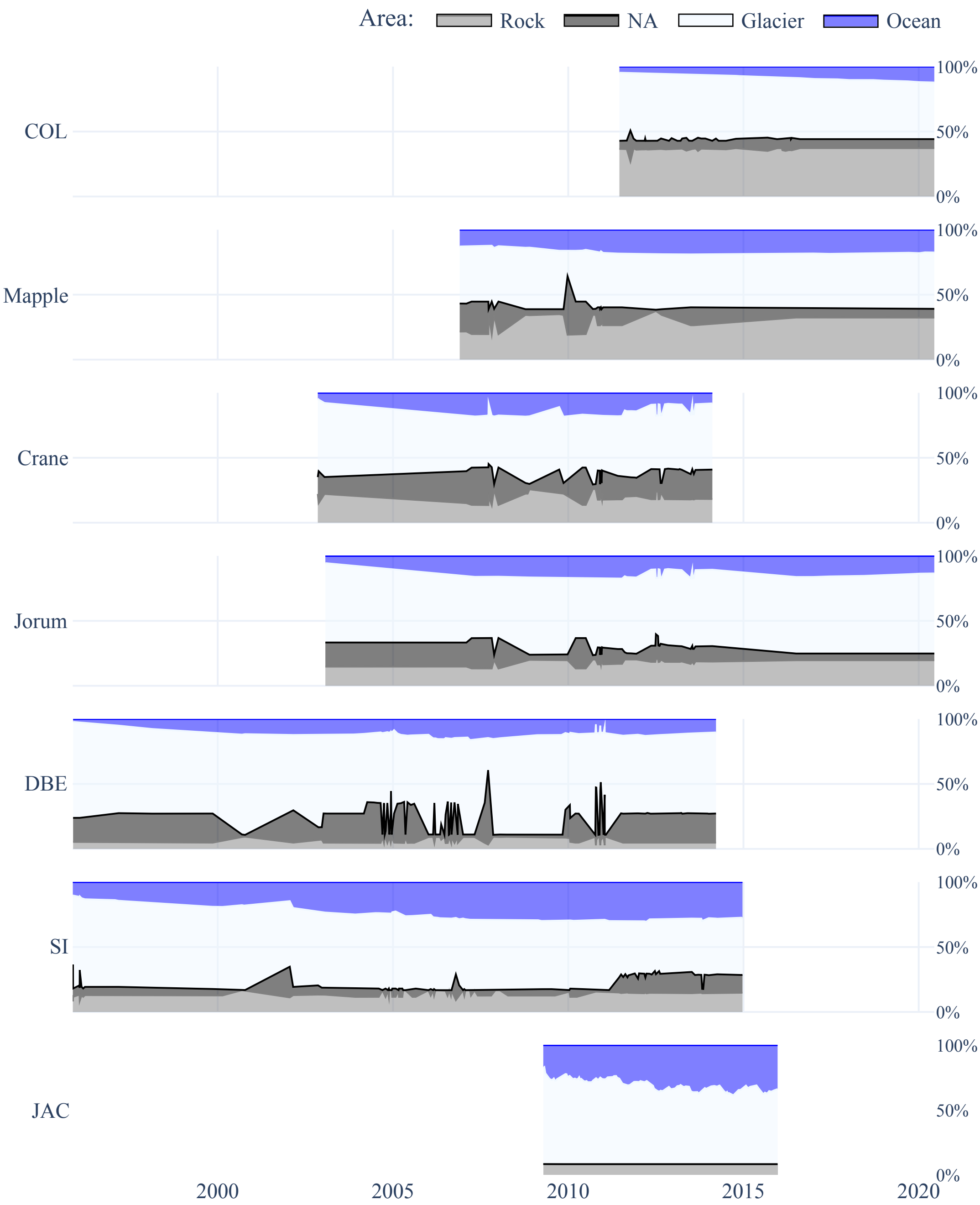


**Figure C1.** Changes of the class distributions in the zone labels of each glacier over time. The ocean is coloured in blue, the glacier area in white, the rock in grey and the area with no available information (NA) in black. There is no radar reflection in the NA area due to terrain elevation causing shadows or due to limited coverage by the radar swath.

## Code and data availability

The code is fundamentally based on the nnU-Net by Isensee et al. (2021), which is also publicly available (https://github.com/MIC-DKFZ/nnUNet). We modified the neural network architecture and corresponding pre- and post-processing. The modified version of the nnU-Net and our evaluation and visualization scripts is available ~~(https://gitlab.cs.fau.de/ho11laqe/nnunet_glacer.git)~~ on GitHub (https://github.com/ho11laqe/nnUNet_calvingfront_detection) and a Demo-version is provided on HuggingFace (https://huggingface.co/spaces/ho11laqe/nnUNet_calvingfront_detection). The calving front predictions of the test set are available on Zenodo (https://zenodo.org/record/7915740#.ZFuP6NIzbUA) as well as the pre-trained model (https://zenodo.org/record/7837300#.ZFuQAdIzbUA). The dataset is provided by Gourmelon et al. (2022) C18 (1.5, 2.3) (https://doi.org/10.1594/PANGAEA.940950) .

## Author contribution

OH and NG were responsible for conceptualizing the study. OH developed and implemented the methodology and software for the experiments. NG and TS provided the dataset and a benchmark score. NG and VC supervised the study and reviewed and edited the manuscript. JF acquired financial support. OH created visualizations and prepared the manuscript with contributions from all co-authors. JF, VC, and NG reviewed and edited the manuscript.

## Competing interests

The authors declare that they have no conflict of interest.

## Acknowledgments

J. J. Fürst has received funding from the European Union’s Horizon 2020 research and innovation programme via the European Research Council (ERC) as a Starting Grant (StG) under grant agreement No. 948290. T. Seehaus was supported by the ESA Living Planet Fellowship project “MIT-AP”. We acknowledge the free provision of the SAR data via various proposals from ESA, ASF, and DLR. The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High-Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU). The hardware is funded by the German Research Foundation (DFG). We acknowledge financial support by Deutsche Forschungsgemeinschaft and Friedrich-Alexander-Universität Erlangen-Nürnberg within the funding programme “Open Access Publication Funding”. This research has been supported by the Friedrich-Alexander-Universität Erlangen-Nürnberg (Emerging Fields Initiative “Tapping the Potential of Earth Observation”) and the STAEDTLER Foundation. The authors thank the Elitenetzwerk Bayern for financial support under the International Doctorate Program “Measuring and modelling mountain glaciers and ice caps in a changing climate (M$^3$OCCA)”.

## References

Abolvardi, A. A., Hamey, L., and Ho-Shon, K.: UNET-Based Multi-Task Architecture for Brain Lesion Segmentation, in: Digital Image Computing: Techniques and Applications (DICTA), pp. 1–7, https://doi.org/10.1109/DICTA51227.2020.9363397, 2020.

Amundson, J. M., Fahnestock, M., Truffer, M., Brown, J., Lüthi, M. P., and Motyka, R. J.: Ice mélange dynamics and implications for terminus stability, Jakobshavn Isbrse, Greenland, Journal of Geophysical Research: Earth Surface, 115, https://doi.org/10.1029/2009JF001405, 2010.

Amyar, A., Modzelewski, R., Li, H., and Ruan, S.: Multi-task deep learning based CT imaging analysis for COVID-19 pneumonia: Classification and segmentation, Computers in Biology and Medicine, 126, 104 037, https://doi.org/10.1016/j.compbiomed.2020.104037, 2020.

Baumhoer, C. A., Dietz, A. J., Dech, S., and Kuenzer, C.: Remote sensing of antarctic glacier and ice-shelf front dynamics-a review, Remote Sensing, 10, 1445, https://doi.org/10.3390/rs10091445, 2018.

Baumhoer, C. A., Dietz, A. J., Kneisel, C., and Kuenzer, C.: Automated Extraction of Antarctic Glacier and Ice Shelf Fronts from Sentinel-1 Imagery Using Deep Learning, Remote Sensing, 11, 2529, https://doi.org/10.3390/rs11212529, 2019.

Baumhoer, C. A., Dietz, A. J., Kneisel, C., Paeth, H., and Kuenzer, C.: Environmental drivers of circum-Antarctic glacier and ice shelf front retreat over the last two decades, The Cryosphere, 15, 2357–2381, https://doi.org/10.5194/tc-15-2357-2021, 2021.

Baumhoer, C. A., Dietz, A. J., Heidler, K., and Kuenzer, C.: IceLines – A new data set of Antarctic ice shelf front positions, Scientific Data, 10, 138, https://doi.org/10.1038/s41597-023-02045-x, 2023.

Beer, C., Biebow, N., Braun, M., Döring, N., Gaedicke, C., Gutt, J., Hagen, W., Hauck, J., Heinemann, G., Herata, H., et al.: Forschungsagenda Polarregionen im Wandel, p. 79, Bundesministerium für Bildung und Forschung (BMBF), Germany, 2021.

Bischke, B., Helber, P., Folz, J., Borth, D., and Dengel, A.: Multi-Task Learning for Segmentation of Building Footprints with Deep Neural Networks, in: International Conference on Image Processing (ICIP), pp. 1480–1484, IEEE, Taipei, https://doi.org/10.1109/ICIP.2019.8803050, 2019.

Carr, J. R., Stokes, C., and Vieli, A.: Recent retreat of major outlet glaciers on Novaya Zemlya, Russian Arctic, influenced by fjord geometry and sea-ice conditions, Journal of Glaciology, 60, 155–170, https://doi.org/10.3189/2014JoG13J122, 2014.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H.: Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation, in: European conference on computer vision (ECCV), edited by Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y., pp. 833–851, Springer International Publishing, Cham, https://doi.org/https://doi.org/10.1007/978-3-030-01234-2_49, 2018.

Chen, S., Bortsova, G., García-Uceda Juárez, A., van Tulder, G., and de Bruijne, M.: Multi-task Attention-Based Semi-supervised Learning for Medical Image Segmentation, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, edited by Shen, D., Liu, T., Peters, T. M., Staib, L. H., Essert, C., Zhou, S., Yap, P.-T., and Khan, A., Lecture Notes in Computer Science, pp. 457–465, Springer International Publishing, Cham, https://doi.org/10.1007/978-3-030-32248-9_51, 2019.

Cheng, D., Hayes, W., Larour, E., Mohajerani, Y., Wood, M., Velicogna, I., and Rignot, E.: Calving Front Machine (CALFIN): glacial termini dataset and automated deep learning extraction method for Greenland, 1972–2019, The Cryosphere, 15, 1663–1675, https://doi.org/10.5194/tc-15-1663-2021, 2021.

Cook, A. J., Murray, T., Luckman, A., Vaughan, D. G., and Barrand, N. E.: A new 100-m Digital Elevation Model of the Antarctic Peninsula derived from ASTER Global DEM: methods and accuracy assessment, Earth System Science Data, 4, 129–142, https://doi.org/10.5194/essd-4-129-2012, 2012.

Davari, A., Baller, C., Seehaus, T., Braun, M., Maier, A., and Christlein, V.: Pixel-wise Distance Regression for Glacier Calving Front Detection and Segmentation, IEEE Transactions on Geoscience and Remote Sensing, 60, 1–10, https://doi.org/10.1109/TGRS.2022.3158591, 2022.

Friedl, P., Seehaus, T. C., Wendt, A., Braun, M. H., and Höppner, K.: Recent dynamic changes on Fleming Glacier after the disintegration of Wordie Ice Shelf, Antarctic Peninsula, The Cryosphere, 12, 1347–1365, https://doi.org/10.5194/tc-12-1347-2018, 2018.

Gourmelon, N., Seehaus, T., Braun, M., Maier, A., and Christlein, V.: Calving fronts and where to find them: a benchmark dataset and methodology for automatic glacier calving front extraction from synthetic aperture radar imagery, Earth System Science Data, 14, 4287–4313, https://doi.org/10.5194/essd-14-4287-2022, 2022.

Gourmelon, N., Seehaus, T., Braun, M. H., Maier, A., and Christlein, V.: CaFFe (CAlving Fronts and where to Find thEm: a benchmark dataset and methodology for automatic glacier calving front extraction from sar imagery), https://doi.org/10.1594/PANGAEA.940950, 2022.

Hartmann, A., Davari, A., Seehaus, T., Braun, M., Maier, A., and Christlein, V.: Bayesian U-Net for Segmenting Glaciers in Sar Imagery, IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 41, 3479–3482, https://doi.org/10.1109/IGARSS47720.2021.9554292, iSSN: 2153-7003, 2021.

He, K., Lian, C., Zhang, B., Zhang, X., Cao, X., Nie, D., Gao, Y., Zhang, J., and Shen, D.: HF-UNet: Learning Hierarchically Inter-Task Relevance in Multi-Task U-Net for Accurate Prostate Segmentation in CT Images, IEEE transactions on medical imaging, 40, 2118–2128, https://doi.org/10.1109/TMI.2021.3072956, 2021.

Heidler, K., Mou, L., Baumhoer, C., Dietz, A., and Zhu, X. X.: HED-UNet: Combined Segmentation and Edge Detection for Monitoring the Antarctic Coastline, IEEE Transactions on Geoscience and Remote Sensing, pp. 1–14, https://doi.org/10.1109/TGRS.2021.3064606, 2021.

Heller, N., Isensee, F., Maier-Hein, K. H., Hou, X., Xie, C., Li, F., Nan, Y., Mu, G., Lin, Z., Han, M., Yao, G., Gao, Y., Zhang, Y., Wang, Y., Hou, F., Yang, J., Xiong, G., Tian, J., Zhong, C., Ma, J., Rickman, J., Dean, J., Stai, B., Tejpaul, R., Oestreich, M., Blake, P., Kaluzniak, H., Raza, S., Rosenberg, J., Moore, K., Walczak, E., Rengel, Z., Edgerton, Z., Vasdev, R., Peterson, M., McSweeney, S., Peterson, S., Kalapara, A., Sathianathen, N., Papanikolopoulos, N., and Weight, C.: The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KiTS19 challenge, Medical Image Analysis, 67, 101 821, https://doi.org/https://doi.org/10.1016/j.media.2020.101821, 2021.

Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., and Maier-Hein, K. H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation, Nature Methods, 18, 203–211, https://doi.org/10.1038/s41592-020-01008-z, 2021.

Jang, H.-J. and Cho, K.-O.: Applications of deep learning for the analysis of medical data, Archives of pharmacal research, 42, 492–504, https://doi.org/https://doi.org/10.1007/s12272-019-01162-9, 2019.

Kholiavchenko, M., Sirazitdinov, I., Kubrak, K., Badrutdinova, R., Kuleev, R., Yuan, Y., Vrtovec, T., and Ibragimov, B.: Contour-aware multi-label chest X-ray organ segmentation, International Journal of Computer Assisted Radiology and Surgery, 15, 425–436, https://doi.org/10.1007/s11548-019-02115-9, 2020.

Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980, 2014.

Kneib-Walter, A., Lüthi, M. P., Moreau, L., and Vieli, A.: Drivers of Recurring Seasonal Cycle of Glacier Calving Styles and Patterns, Frontiers in Earth Science, 9, https://doi.org/10.3389/feart.2021.667717, 2021.

Li, X., Sun, X., Meng, Y., Liang, J., Wu, F., and Li, J.: Dice loss for data-imbalanced NLP tasks, arXiv preprint arXiv:1911.02855, 2019a.

Li, X., Wang, Y., Tang, Q., Fan, Z., and Yu, J.: Dual U-Net for the Segmentation of Overlapping Glioma Nuclei, IEEE Access, 7, 84 040–84 052, https://doi.org/10.1109/ACCESS.2019.2924744, 2019b.

Loebel, E., Scheinert, M., Horwath, M., Heidler, K., Christmann, J., Phan, L. D., Humbert, A., and Zhu, X. X.: Extracting glacier calving fronts by deep learning: the benefit of multi-spectral, topographic and textural input features, IEEE Transactions on Geoscience and Remote Sensing, 60, 1–12, https://doi.org/10.1109/TGRS.2022.3208454, 2022.

Marochov, M., Stokes, C. R., and Carbonneau, P. E.: Image Classification of Marine-Terminating Outlet Glaciers usingDeep Learning Methods, preprint, Glaciers/Remote Sensing, https://doi.org/10.5194/tc-2020-310, 2020.

McNabb, R. W., Hock, R., and Huss, M.: Variations in Alaska tidewater glacier frontal ablation, 1985–2013, Journal of Geophysical Research: Earth Surface, 120, 120–136, https://doi.org/https://doi.org/10.1002/2014 JF003276, 2015.

Minowa, M., Schaefer, M., Sugiyama, S., Sakakibara, D., and Skvarca, P.: Frontal ablation and mass loss of the Patagonian icefields, Earth and Planetary Science Letters, 561, 116 811, https://doi.org/https://doi.org/10.1016/j.epsl.2021.116811, 2021.

Mohajerani, Y., Wood, M., Velicogna, I., and Rignot, E.: Detection of Glacier Calving Margins with Convolutional Neural Networks: A Case Study, Remote Sensing, 11, 74, https://doi.org/10.3390/rs11010074, 2019.

Periyasamy, M., Davari, A., Seehaus, T., Braun, M., Maier, A., and Christlein, V.: How to Get the Most Out of U-Net for Glacier Calving Front Segmentation, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 15, 1712–1723, https://doi.org/10.1109/JSTARS.2022.3148033, 2022.

Recinos, B., Maussion, F., Rothenpieler, T., and Marzeion, B.: Impact of frontal ablation on the ice thickness estimation of marine-terminating glaciers in Alaska, The Cryosphere, 13, 2657–2672, https://doi.org/10.5194/tc-13-2657-2019, 2019.

Recinos, B., Maussion, F., Noël, B., Möller, M., and Marzeion, B.: Calibration of a frontal ablation parameterisation applied to Greenland's peripheral calving glaciers, Journal of Glaciology, 67, 1177–1189, https://doi.org/10.1017/jog.2021.63, 2021.

Robel, A. A., Schoof, C., and Tziperman, E.: Persistence and variability of ice-stream grounding lines on retrograde bed slopes, The Cryosphere, 10, 1883–1896, https://doi.org/10.5194/tc-10-1883-2016, 2016.

Ronneberger, O., Fischer, P., and Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, in: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, edited by Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F., vol. 9351, pp. 234–241, Springer International Publishing, Cham, https://doi.org/10.1007/978-3-319-24574-4_28, 2015.

Rott, H., Wuite, J., Rydt, J. D., Gudmundsson, G. H., Floricioiu, D., and Rack, W.: Impact of marine processes on flow dynamics of northern Antarctic Peninsula outlet glaciers, Nature Communications, 11, 2969, https://doi.org/10.1038/s41467-020-16658-y, 2020.

Shepherd, A., Ivins, E., Rignot, E., Smith, B., Broeke, M. V. D., Velicogna, I., Whitehouse, P., Briggs, K., Joughin, I., Krinner, G., Nowicki, S., Payne, T., Scambos, T., Schlegel, N., Geruo, A., Agosta, C., Ahlstrøm, A., Babonis, G., Barletta, V., Blazquez, A., Bonin, J., Csatho, B., Cullather, R., Felikson, D., Fettweis, X., Forsberg, R., Gallee, H., Gardner, A., Gilbert, L., Groh, A., Gunter, B., Hanna, E., Harig, C., Helm, V., Horvath, A., Horwath, M., Khan, S., Kjeldsen, K. K., Konrad, H., Langen, P., Lecavalier, B., Loomis, B., Luthcke, S., McMillan, M., Melini, D., Mernild, S., Mohajerani, Y., Moore, P., Mouginot, J., Moyano, G., Muir, A., Nagler, T., Nield, G., Nilsson, J., Noel, B., Otosaka, I., Pattle, M. E., Peltier, W. R., Pie, N., Rietbroek, R., Rott, H., Sandberg-Sørensen, L., Sasgen, I., Save, H., Scheuchl, B., Schrama, E., Schröder, L., Seo, K. W., Simonsen, S., Slater, T., Spada, G., Sutterley, T., Talpe, M., Tarasov, L., Berg, W. J. V. D., Wal, W. V. D., Wessem, M. V., Vishwakarma, B. D., Wiese, D., and Wouters, B.: Mass balance of the Antarctic Ice Sheet from 1992 to 2017, Nature, 558, 219–222, https://doi.org/10.1038/s41586-018-0179-y, 2018.

Smith, L. N.: Cyclical Learning Rates for Training Neural Networks, in: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 464–472, https://doi.org/10.1109/WACV.2017.58, 2017.

Straneo, F., Heimbach, P., Sergienko, O., Hamilton, G., Catania, G., Griffies, S., Hallberg, R., Jenkins, A., Joughin, I., Motyka, R., Pfeffer, W. T., Price, S. F., Rignot, E., Scambos, T., Truffer, M., and Vieli, A.: Challenges to Understanding the Dynamic Response of Greenland's Marine Terminating Glaciers to Oceanic and Atmospheric Forcing, Bulletin of the American Meteorological Society, 94, 1131–1144, https://doi.org/10.1175/BAMS-D-12-00100.1, 2013.

Zhang, E., Liu, L., and Huang, L.: Automatically delineating the calving front of Jakobshavn Isbræ from multitemporal TerraSAR-X images: a deep learning approach, The Cryosphere, 13, 1729–1741, https://doi.org/10.5194/tc-13-1729-2019, 2019.

Zhang, E., Liu, L., Huang, L., and Ng, K. S.: An automated, generalized, deep-learning-based method for delineating the calving fronts of Greenland glaciers from multi-sensor remote sensing imagery, Remote Sensing of Environment, 254, 112 265, https://doi.org/10.1016/j.rse.2020.112265, 2021.