Technische Universität München
TUM School of Engineering and Design

Interpretation of Structures in Polar Regions with Deep Learning Methods

Konrad Marten Harald Heidler

Vollständiger Abdruck der von der TUM School of Engineering and Design der Technischen Universität München zur Erlangung eines

Doktors der Ingenieurwissenschaften (Dr.-Ing.)

genehmigten Dissertation.

Vorsitz:

Prof. Dr. Jonathan Bamber

Prüfende der Dissertation:

1. Prof. Dr.-Ing. habil. Xiaoxiang Zhu
2. Prof. Dr.-Ing. habil. Richard Hans Georg Bamler
3. Prof. Sébastien Lefèvre

Die Dissertation wurde am 22.01.2024 bei der Technischen Universität München eingereicht und durch die TUM School of Engineering and Design am 06.05.2024 angenommen.

Abstract

Global climate change is rapidly transforming the polar regions. Remote sensing offers great potential for monitoring the developments in these often inaccessible areas. However, the amount of data is too large for manual analysis. Therefore, in an effort to support monitoring the changes in these regions, this dissertation develops deep learning methods for the remote sensing analysis of targets in these regions. More specifically, the developed models can automatically map the calving fronts of marine-terminating glaciers and detect permafrost disturbances in the form of retrogressive thaw slumps (RTS). The fundamental research questions motivating this dissertation are:

1. How can domain knowledge about polar regions be encoded into deep learning models for remote sensing in these regions?
2. Are there more efficient ways of encoding such mapping tasks into deep learning tasks than the standard approaches used in computer vision like semantic segmentation?
3. How can deep learning models generalise from limited labels to the entirety of the polar regions in a data-efficient way?
4. Will automatically derived observations allow polar science to better understand the developments in polar regions?

In an effort to answer these questions, this dissertation makes the following scientific contributions:

1. A model for mapping calving fronts in Antarctica is developed based on observations in the behaviour of human annotators. The resulting HED-UNet model combines semantic segmentation and edge detection and works on multiple resolution levels [1].
2. Questioning the representation of calving fronts through pixel-wise predictions altogether, a second model is developed. The COBRA model maps calving fronts in Greenland by directly predicting the desired contour lines instead of taking pixel-wise predictions as a proxy [2].
3. In a first downstream study, the COBRA model is applied for calving front detection in Svalbard. Thanks to the automated analysis, it was possible to derive a dataset of more than 100,000 calving front traces. This dataset allows for a better understanding of glacial processes such as the behaviour of surge-type glaciers [3].

4. The feasibility of applying deep learning methodology for the mapping of retrogressive thaw slumps in permafrost regions is established. By evaluating multiple deep learning architectures with regional cross-validation across the Arctic, spatial generalisation is identified as the main challenge [4].

5. In order to address the challenge of spatial generalisation in permafrost disturbance mapping, a data-efficient training routine is proposed. The PixelDINO approach is a method for semi-supervised learning, combining labelled data with unlabelled imagery. In this way, models are trained to generalise well across multiple regions in the Arctic [5].

Zusammenfassung

Der globale Klimawandel hat massive Auswirkungen auf die Polarregionen. Für die Beobachtung der Entwicklungen in diesen oft unzugänglichen Regionen stellt die Fernerkundung ein wertvolles Werkzeug dar. Jedoch ist die Menge an Daten zu groß für manuelle Analysen. Um die Analysen dieser Prozesse mit automatisierten Methoden zu unterstützen, entwickelt diese Dissertation Deep Learning Methoden für die Fernerkundung von bestimmten Objekten in den Polarregionen. Die entwickelten Modelle können Gletscherkalbungsfronten kartieren, sowie Störungen des Permafrostbodens in der Form von Retrogressiven Taurutschungen detektieren. Die motivierenden Forschungsfragen sind hierbei die folgenden:

1. Wie kann Anwendungswissen über die Polarregionen in Deep Learning Modelle für die Fernerkundung eingebacht werden?

2. Gibt es effizientere Wege, diese Kartierungsaufgaben in Deep Learning Modellen zu kodieren als die geläufigen Computer Vision Methoden, wie semantische Segmentierung?

3. Wie kann es Deep Learning Modellen ermöglicht werden, über große Areale zu generalisieren, ohne den Aufwand für die Erstellung von Trainingsdaten beträchtlich zu steigern?

4. Inwieweit können die automatisch abgeleiteten Vorhersagen der Polarforschung helfen, Prozesse in den Polarregionen besser zu verstehen?

In dem Bestreben, diese Fragen zu beantworten, liefert diese Dissertation die folgenden wissenschaftlichen Beiträge:

1. Basierend auf Beobachtungen zum Verhalten von menschlichen Annotatoren wird ein Modell zur Kartierung von Kalbungsfronten in der Antarktis entwickelt. Das resultierende HED-UNet Modell kombiniert dafür semantische Segmentierung mit Kantendetektion und verarbeitet die Bilddaten auf mehreren Auflösungsebenen [1].

2. Die Repräsentation von Kalbungsfronten durch pixelweise Vorhersagen wird infrage gestellt. Das zweite entwickelte Modell, COBRA, kartiert Kalbungsfronten direkt in der Form von Polygonzügen. Somit wird der Umweg über pixelweise Masken vermieden [2].

3. In einer ersten Anwendungsstudie wird das COBRA Modell auf Kalbungsfronten in Spitzbergen angewandt. Dank der automatisierten Analyse war es möglich, einen Datensatz von über 100.000 Kalbungsfronten zu generieren. Dieser Datensatz gibt neue Einblicke in das Verständnis von Gletscherprozessen, wie zum Beispiel das Verhalten von Surge-Gletschern [3].

4. Eine Machbarkeitsstudie demonstriert das Potential von Deep Learning für die Kartierung von Retrogressiven Taurutschungen in Permafrostregionen. Durch die Evaluation von verschiedenen Deep Learning Architekturen und regionaler Kreuzvalidierung wird die räumliche Generalisierung als zentrale Herausvorderung identifiziert [4].

5. Um die räumliche Generalisierung der Modelle zu verbessern wird eine dateneffiziente Trainingsprozedur vorgestellt. Der PixelDINO-Ansatz ist eine Methode für Semi-Supervised Learning, wobei existierende annotierte Trainingsdaten mit nicht annotierten Satellitenbildern kombiniert werden. Auf diese Weise ist es möglich, Modelle zu trainieren, die Vorhersagen von hoher Qualität für diverse Regionen der Arktis liefern [5].

Acknowledgments

This dissertation was a huge effort that I could never have hoped to complete without the support of others. During my time as a doctoral researcher, I have had the pleasure of meeting countless people who supported my journey. I would like to take this opportunity to say thanks to all of you, I could not have done it without you.

First, I want to thank my supervisor, Xiaoxiang Zhu, for guiding me all the way along this journey. Her encouraging support throughout this dissertation and her trust in my research ideas were indispensable foundations for my endeavours. I would also like to thank Richard Bamler for his continued interest and encouraging comments on my work, and for agreeing to act as an examiner for my dissertation. I express my gratitude to Sébastien Lèfevre, not only for acting as an external examinar for my dissertation, but also for insightful discussions about unconventional deep learning approaches, and for encouraging me to stay persistent in a project that I had nearly written off. I extend my thanks to Jonathan Bamber for sharing his expertise and insights on the topic of glaciers, and for kindly taking over chairmanship of my doctoral defense.

Next, I would like to thank the colleagues who supported my work. When I started my Ph.D. journey at the German Aerospace Center just in time for the first COVID-19 lockdown, Anja Rösel, Eike Jens Hoffmann and Yingjie Schreiber-Liu helped me to get started as smoothly as possible in those uncertain times. I thank Lichao Mou, my mentor, for many stimulating discussions on various deep learning topics, and for teaching me how to structure my everyday research. Special thanks go to all my collaboration partners, who worked together with me towards a shared vision of improving polar remote sensing with deep learning. In particular, I would like to highlight Ingmar Nitze, Erik Loebel, Celia Baumhoer, and Tian Li, who all spent many hours answering my (at times completely ignorant) questions about cold regions and discussing ideas. Moving to TUM later on, I could always rely on the amazing support of the science management team, Julia Kollofrath, Simon Schneider, Marcus Langejahn, and Vicky Karasmanaki. No matter what I needed, or when, they made it happen.

During my time at DLR and TUM, many colleagues enriched my life as a researcher. A special thanks to my office mate Adrian, who was always happy to discuss the joys and hardships of our Ph.D. projects, and gave valueable feedback on my ideas. I would also like to thank Nassim for countless enjoyable discussions, be they on scientific topics like self-supervised learning and large language models, or on ridiculous topics like our favourite Pokémon. I am grateful to Nikolai, Ivica, Cédric, Nils, Viola, Paul, Matthias, Tabea, Saskia, Constantin, Teo, Christoph, and so many others for enjoyable coffee breaks and conversations that made my Ph.D. experience so much more delightful. I thank Mohsin and Codruț for the fun evenings we spent climbing, allowing us to forget about the stress that a Ph.D. projects can entail at times. Together with my

MDSI buddies Martin, Florian and Kathrin, I spent countless joyful days co-working at the Munich Data Science Institute. Thanks for the inspiring discussions about the bleeding-edge topics in deep learning research.

Last but not least, I want to thank my parents, Imke and Harald, and my sister, Sonja, for always believing in me, even in times where I started doubting myself. Above all, my heartfelt thanks go to my fiancée Katharina, whose unwavering support and countless pep talks have made all the difference. I am truly grateful to know you by my side not just on this adventure, but for many more to come.

Konrad Heidler
Munich, January 2024

Contents

List of Figures

1 Introduction

The regions near the Earth's poles, the Arctic and Antarctica, are characterised by frigid climates, giving rise to unique ecosystems found nowhere else on Earth. Icy landforms like ice sheets or permafrost comprise large parts of these regions. Moreover, the polar regions are home to iconic species like penguins and polar bears. However, these regions are changing at alarming rates in the face of global climate change. In fact, out of all regions on Earth, both Antarctica and the Arctic are among the regions warming most rapidly. They are estimated to currently be warming at 2–4 times the global average warming rate, a phenomenon called polar amplification [6], [7]. Since the polar regions are covered primarily by icy landforms, these warming temperatures manifest themselves in the form of melting processes. This melting has massive implications for local geophysics and ecosystems. Glaciers in Greenland and Antarctica are mostly retreating, often at accelerating rates [8], [9]. Depending on the emission scenarios, the Greenland and West Antarctic ice sheets might even disappear completely [10]. Arctic sea ice has consistently been declining, especially in the 1980s and since 2010 [10], [11]. Similarly, ground temperatures are increasing in many permafrost regions, causing previously frozen soil to thaw rapidly [12]. Most of these changes are very likely anthropogenic in nature, meaning that they are caused by the effect humans have on the global climate [10].

These developments in the polar regions may seem to be locally constrained at first glance. One might easily dismiss them as irrelevant, especially since the polar regions are sparsely populated. However, we cannot ignore these developments, as the global climate system is tightly interconnected. Even small changes in one region can lead to significant changes in other regions [13], [14]. Ultimately, warming processes in the polar regions are not only of regional importance. On the contrary, they are likely to impact multiple global climate systems [14]. For instance, melting glaciers in Greenland and Antarctica are a significant contributor to the rise of global sea levels, with measurements indicating a contribution of around 7.6 mm from the Antarctic Ice Sheet for the period from 1992 to 2017 [15] and around 10.8 mm from the Greenland Ice Sheet for the period from 1992 to 2018 [16]. The ice sheets are tightly coupled to the global climate system not only through sea level rise, but also through transport mechanisms like ocean currents and air fluxes [17], [18]. In the Arctic Ocean, sea temperatures and salinity are rising, while sea ice is declining, a process called *Atlantification* [19]. In turn, important ocean currents in the Arctic Ocean are changing their behaviour, which could have global repercussions [20].

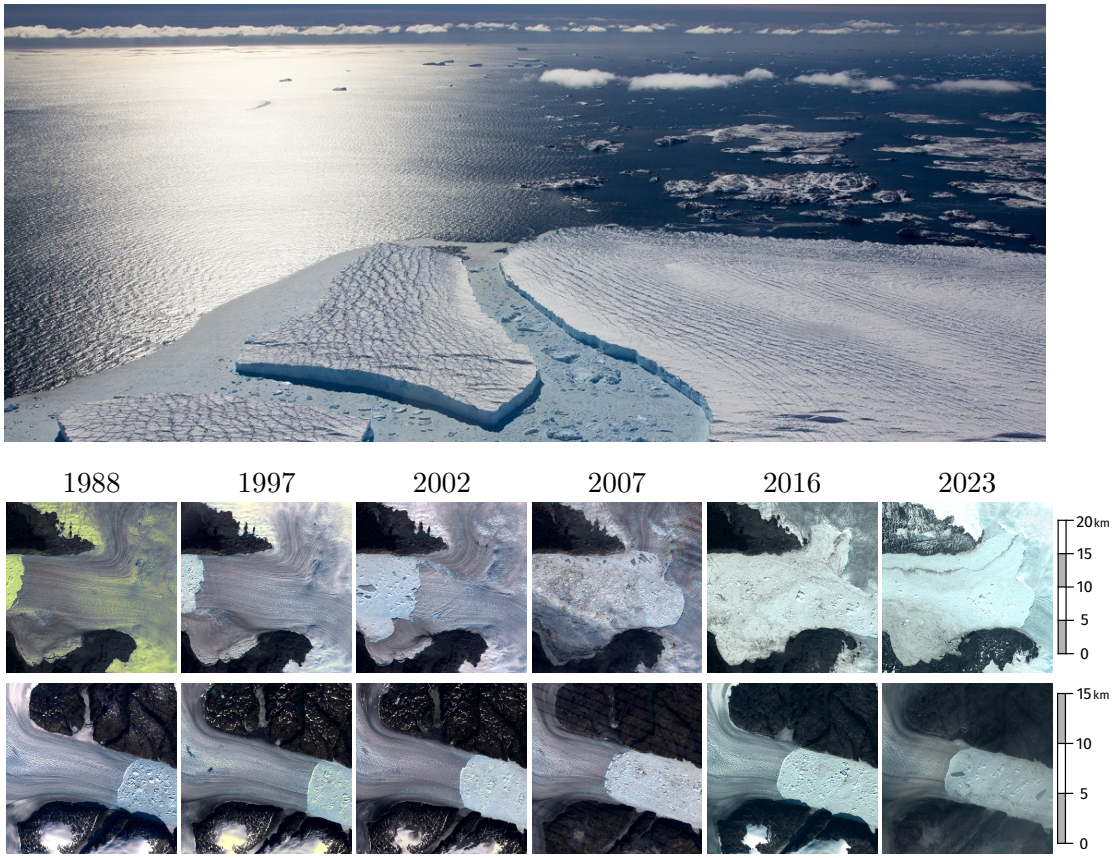


Figure 1.1: Top: Calving event of Sørsdal Glacier in Antarctica, Image by David Gwyther [21]. Bottom: Timeseries of retreating glaciers in Greenland. First Row: Jakobshavn Isbræ in Western Greenland, retreating from west (left) to east (right). Second Row: Helheim Glacier in Eastern Greenland, retreating from east (right) to west (left). Satellite imagery from the NASA Landsat 5, 7, 8, and 9 missions, processed using the Google Earth Engine [22].

Perhaps the most prominent features of the polar regions are the massive ice shields in Greenland and Antarctica. In light of extensive warming in these regions, the glaciers fed by these ice sheets are already losing mass at accelerating rates, as shown in figure 1.1, which causes global sea levels to rise [10], [15], [16]. Experts believe these ice shields will lose substantial amounts of their ice mass by the end of the century [10]. Research suggests that these processes are part of intricate feedback loops where increased glacier melt increases the interactions between warmer ocean water and the glacial ice [13]. Further, glacial retreat lowers the albedo of the Earth's surface so that more incoming radiative energy from the sun will be absorbed [23]. Both of these effects, in turn, accelerate the melting process of the glaciers. This self-reinforcement of glacial melt in the polar regions is one reason for significant concerns regarding so-called

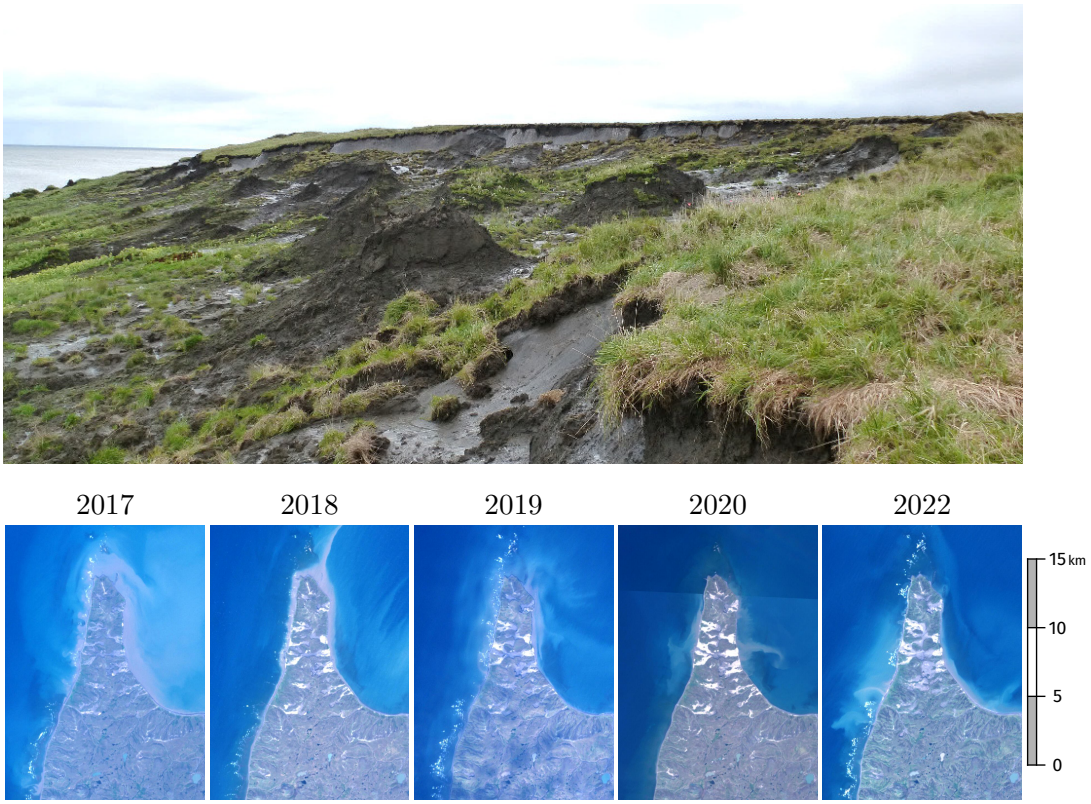


Figure 1.2: Top: Coastal retrogressive thaw slump on the Bykovsky Peninsula in northern Siberia. Image by Ingmar Nitze [24]. Bottom: Timeseries of rapid permafrost decay through growing RTS (bright, beige features) at Sukhoy Nos on the Novaya Zemlya archipelago in the Russian Arctic. Satellite imagery from the ESA Sentinel-2 mission, downloaded and processed using Google Earth Engine [22].

tipping points, which are thresholds that, once crossed, will cause abrupt changes in the global climate system, which will not be able to be reversed [23].

Another process driven by the warming of the Arctic is the thawing of permafrost. The term *permafrost* denotes soil that remains at 0 °C or below for at least two consecutive years [25]. Similarly to the ice sheets, permafrost makes up a significant fraction of the Earth's surface. Accurately quantifying the area underlain by permafrost is challenging because it is a sub-surface phenomenon. Assessments agree, however, that at least 10% of the Earth's land surface is currently underlain by permafrost [26]. Warming in these regions causes this previously frozen soil to thaw gradually, causing massive repercussions for both local ecosystems and the global climate. Thawed permafrost soil has entirely different physical properties from intact permafrost [27]. Therefore, these thawing processes can change entire landforms, as seen in figure 1.2. For example, parts of the ground can subside, posing threats to local infrastructure like roads, houses, or pipelines [28]. Further, thawing permafrost is known to release previously stored organic carbon in the atmosphere, mainly in the form of the greenhouse gas methane [29].

Climate researchers believe this mechanism will further drive climate change through a so-called feedback loop [29]. As permafrost is a subsurface phenomenon, these thawing processes are challenging to understand on a global scale. Global model estimates of permafrost parameters like the ground temperature or the active permafrost layer thickness deviate considerably from the ground truth in many places [30], [31].

While scientists agree that the effects of rapid warming in the polar regions are potentially catastrophic [10], it is not easy to quantify them precisely, as the underlying processes are complex to observe and model. While some phenomena are well understood qualitatively, they remain challenging to model quantitatively, especially when it comes to forecasting future developments. The process of icebergs calving from glaciers, for example, has been extensively studied through numerical models [32]–[34]. Still, each model has shortcomings, and none can currently predict iceberg calving with reliable accuracy [35], [36]. More real-world data measurements are needed to gain a better understanding of such dynamics and develop more accurate models of these processes. These measurements must have both a high spatial resolution and a high temporal frequency so that the underlying processes can be reconstructed and understood in sufficient detail.

Highly accurate and frequent data about cryospheric processes are needed. The actual acquisition of such data, however, is not an easy task. Traditionally, studies on the polar regions are carried out in the form of expeditions. A group of researchers travels to a specific polar region and takes measurements in the field. The remote location and difficult accessibility of large parts of the Arctic impose significant restrictions on the coverage of such studies. Expeditions can only reach a small number of selected study sites, while most of the Arctic remains uncovered by such expeditions. Such accessibility issues are even more pressing for Antarctic expeditions. While local expeditions provide deep insight into specific local geophysics and ecosystems, they are costly and only able to visit a few locations at a time. Therefore, it is challenging to directly measure or observe trends from field studies alone. Especially on a pan-Arctic or pan-Antarctic, in-situ data is scarce [37].

Thanks to the remarkable recent technological progress of Earth observation satellites and data processing, we can virtually explore and observe any place on Earth today without physically travelling there. Vast amounts of satellite data are acquired daily, allowing for insights with unprecedented resolution and accuracy even for secluded regions that might otherwise go unobserved. These possibilities are of high practical interest for the polar regions, as these regions are, in large parts, inaccessible even to scientific expeditions. Remote sensing has, therefore, become an invaluable tool for observing structures and processes in these regions [38], [39]. Especially in light of global climate change causing unprecedented developments that are ever-increasing in speed and magnitude, satellite imagery provides a consistent monitoring possibility for the Arctic and Antarctica. In order to gain a deeper understanding of these changes and the underlying processes, remote sensing can help by allowing for large-scale analysis of these phenomena.

Scaling up such monitoring systems to the entirety of the Arctic or Antarctica requires processing vast amounts of data. Therefore, manual analysis is not an option.

For such analyses, recent data science methods can automate many tedious tasks, like mapping objects of interest, classifying features on the ground, or tracking changes. Once such an automated algorithm has been designed and trained, it is relatively simple to scale up the analysis by increasing the amount of computational resources, which is more economical than employing hundreds of manual annotators. This premise serves as the starting point for this dissertation project. The central goal of this thesis is to develop and provide methods based on deep learning algorithms for the automated analysis of phenomena in polar regions. The observations derived using these algorithms can then aid polar scientists in finding patterns, detecting trends, and improving existing numerical models.

1.1 Research Objectives

The main goal of this dissertation is to develop and present automated analysis methods for monitoring polar regions using remote sensing data. In order to achieve this, this dissertation builds on and improves upon successful deep-learning methods from computer vision. Several considerations arise in pursuit of this main goal, which will be outlined in the following paragraphs.

Remote sensing works with imagery quite different from the imagery used in most computer vision research. Differences in imaging resolution, number of spectral bands, and acquisition geometries suggest that analysing remote sensing imagery might require different algorithmic strategies than natural imagery. Therefore, the assumptions behind the design of existing computer vision models may not hold for remote sensing. Even for existing remote sensing models, one might wonder whether a model designed for urban or agricultural remote sensing should be applied to the polar regions without any additional changes. These considerations form the first research question that this thesis aims to answer:

> **Research Question**
>
> Can domain knowledge about polar research be encoded in deep learning models, and if so, how does it improve model predictions compared to existing deep learning approaches for image analysis?

The desired monitoring tasks must be encoded into a task representation well-suited for deep learning. As we will see in chapter 4, there can be multiple ways of doing this, and some may be better suited for the underlying task than others. For instance, mapping tasks in remote sensing are often approached by translating them into the deep learning task of semantic segmentation. Here, the model assigns a class label to each input pixel. For further processing in a geographic information system (GIS) context, the model then needs to extract the desired polygons from the resulting segmentation map in a post-processing step. In search for alternative approaches, the first two manuscripts in this dissertation challenge this approach by exploring different ways of encoding glacier calving front mapping into deep learning tasks.

Research Question

Aside from semantic segmentation, are there other ways of encoding domain-specific tasks, such as calving front detection, that improve prediction performance?

Further, generalisation across different parts of the polar regions deserves careful attention. The models should not only work well for a single study site within the Arctic or Antarctica but must also generalise to the entirety of the polar regions. In this way, large-scale studies of the trends and phenomena mentioned above can be conducted without additional manual image analysis. This is a challenging requirement, especially in the permafrost use case. In order to generalise well, deep learning models need a large amount of representative training data. However, since permafrost landscapes cover large parts of the planet's land surface, they are highly diverse. Therefore, existing datasets only cover small regions within the Arctic limiting the number of available training labels. Chapter 5 will explore ways to improve model generalisation in the setting of limited labels. Semi-supervised learning approaches can use extensive collections of unlabelled imagery to complement the labelled data. The model can then infer structural similarities between the labelled and unlabelled training inputs during training and exploit this to learn to generalise better.

Research Question

How can machine learning models learn efficiently from limited labels in the face of the many diverse landscapes in the Arctic?

Finally, the developed models must be able to analyse remote sensing imagery automatically and extract insights on a large scale. The scientific value of these derived datasets for downstream research glaciology and climate science mainly depends on the data quality. Exemplary studies in Antarctica [9] and Svalbard [3] already employ two of the models proposed in this dissertation in order to derive data about the glaciers in these regions with unprecedented temporal and spatial resolution.

Research Question

To what extent are the data products derived from deep learning models beneficial for polar science?

1.2 Thesis Outline

This section briefly introduces the thesis outline as a guide for the reader. As the presented work constitues a cumulative dissertation, the main scientific contributions

Chapter 2 introduces the necessary concepts from deep learning and remote sensing to understand the remaining chapters. Then, Chapter 3 summarises existing work

on polar remote sensing, highlighting approaches based on traditional remote sensing methods and first studies employing deep learning for these tasks. Finally, it also discusses research works parallel to this dissertation.

Going into the research done during this dissertation project, chapter 4 introduces the first three papers of this dissertation project. Here, we study deep learning for glacier calving front detection. The first paper introduces HED-UNet, a method for combining semantic segmentation and edge detection approaches for more accurate calving front detections. The second paper proposes the COBRA model, which questions the necessity of pixel-wise predictions and instead aims at predicting calving fronts directly as contours. The third paper is a study applying the COBRA model to glaciers in Svalbard on a large scale, demonstrating its usefulness and accuracy in a practical context.

Chapter 5 then discusses the detection of retrogressive thaw slumps in permafrost regions. The first paper introduced in this chapter demonstrates the overall feasibility of deep learning for this task. At the same time, it highlights the need for more sophisticated methods in order to generalise across the Arctic. The second paper proposes the PixelDINO training methodology to address this need. By incorporating unlabelled satellite imagery in a semi-supervised fashion, the models can learn more general and robust features without needing additional labelled data. As we will see, this training procedure can greatly improve the quality of the model predictions.

Finally, chapter 6 concludes the thesis and gives an outlook on future developments in polar remote sensing. It also highlights some open research questions in this area. The appendix contains the publications that make up this cumulative dissertation.

A companion website containing links to the source code repositories and additional material like animations and interactive maps for the publications in this thesis can be found at `https://konrad.heidler.info/dissertation` or by scanning the QR Code below:

2 Theoretical Background

Before going into detail on the current research on polar remote sensing, this chapter aims to familiarise the reader with the background knowledge needed to understand the following chapters. Starting from a short introduction to deep learning as the primary algorithmic technique used in the research projects, we will then move on to how to apply deep learning to remote sensing data, and, finally, explore some of the particularities and challenges of remote sensing in the polar regions.

In order to stay within the scope of this dissertation, we will only visit some concepts in brevity, with references pointing the interested reader to in-depth accounts of the matter at hand.

2.1 Deep Learning: A Primer

Thanks to the large interest in deep learning, many introductory works on this topic are available. Therefore, the following section summarises only the basics necessary to follow the remainder of this dissertation. Readers looking for an in-depth introduction to the foundations of deep learning are referred to Goodfellow *et al.* [40] or Bishop [41]. For an introduction of the topic from a statistical point of view, the textbook by Murphy [42] gives probabilistic intuitions on many of the concepts in deep learning.

Conceptually, deep learning describes a collection of techniques and methods based around deep neural networks (DNNs). Inspired by the connectivity structures of *neurons* in the human brain, they are a particular class of mathematical functions that can be tuned through their *parameters*. DNNs map values from some input vector space $\mathbb{R}^I$ to some output space $\mathbb{R}^O$. Intriguingly, these networks are built from only elementary mathematical operations but are still able to approximate a large class of possible functions up to arbitrary precision [41].

Multi-Layer Perceptrons

As the oldest class of DNNs, Multilayer Perceptrons (MLPs), or feed-forward networks (FFNs) are made up of a sequence of *layers*, which each contain a pre-defined number of artificial neurons. The mapping from one layer to the next is then defined as a mapping

$$x \longmapsto \varphi(Wx + b), \tag{2.1}$$

where W is called the *weight-matrix* and b is called the *bias vector*. φ denotes a *non-linearity* function, needed to approximate non-affine functions. In theory, any non-linear function can be used here, but in practice, simple functions like the positive part of a number, the so-called rectified linear unit (ReLU), are often used. The parameters W and b are called the layer's *parameters*. They determine how the layer transforms its inputs and can be adapted to change the behaviour of the resulting function. An MLP is the functional concatenation of such layers. The collection of trainable parameters of a DNN is often denoted by θ and the resulting mapping for a specific parameter set by $f_\theta(\,\cdot\,)$. The *output layer* is the last layer of an MLP. Earlier layers are called *hidden layers* [40].

While this class of models is rather simplistic from a mathematical point of view, it turns out that such models can approximate a large class of functions. So-called *universal approximation theorems* state that by increasing either the width of the hidden layers [43] or adding more hidden layers [44], one can approximate any continuous function from a compact subset of $\mathbb{R}^I$ to $\mathbb{R}$ with any desired accuracy. This adaptability makes MLPs a potent tool for implementing functions that are hard to define explicitly.

So we have established that DNNs have some impressive capabilities, but how can they be trained to be made helpful for real-world use cases? In machine learning tasks, a typical setting is *supervised learning*, where one is looking to learn an underlying mapping from a set of input-output pairs $\{(x_1, y_1), \dots, (x_N, y_N)\}$. After fixing a neural network architecture, supervised learning is equivalent to finding an optimal set of parameters θ^*, for which $f_{\theta^*}(x_k) \approx y_k$ for all k. For this, some notion of "closeness" of the model output $f_\theta(x_k)$ and the desired output y_k is needed. A *loss function* does precisely that. Given a model output and the true target value, it computes a measure of distance. This distance can then be minimised by means of numerical optimisation. So more formally, given a loss function $\mathcal{L}$, the goal of neural network training is to find θ^* according to

$$\theta^* = \operatorname{argmin}_\theta \frac{1}{N} \sum_{k=0}^{N} \mathcal{L}(f_\theta(x_k), y_k), \tag{2.2}$$

a procedure called *empirical risk minimisation*. The probabilistic intuition behind this is founded in the idea that (x_k, y_k) are independent, identically distributed (i.i.d.) samples drawn from an underlying joint distribution $P(X, Y)$. In this setting, the objective given in equation (2.2) is equivalent to minimising the expected loss for a random sample from this distribution [40].

In practice, some form of gradient descent is most often used to minimise the loss. Following the usual definitions, all components of a neural network are differentiable almost everywhere, meaning that a gradient $\nabla_\theta f_\theta(x)$ of the neural network's outputs with regard to its parameters can be found for a given input x. As the gradient points in the direction of maximum local steepness, updating θ in the negative gradient direction will decrease the loss when taking small enough steps. As datasets tend to be too large to be processed at once, the gradient is usually approximated using a small subset of all samples by randomly sampling indices $B \subset \{1, \dots, N\}$, called a *minibatch*. Then,

the weights are updated according to the gradient estimated from this minibatch:

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta \frac{1}{|B|} \sum_{k \in B} \mathcal{L}(f_\theta(x_k), y_k) \tag{2.3}$$

Repeating this for many steps yields the training procedure of stochastic gradient descent (SGD) [40].

While gradient descent procedures are not guaranteed to converge to the global optimum θ^* in general, optimisation results are largely satisfactory in practice [40]. Modern deep learning optimisers like *Adam* are variations on this basic algorithm which include estimates of higher-order statistics of the gradient [45].

Convolutional Neural Networks

While MLPs have the impressive learning capabilities mentioned earlier, they can become unwieldy for real-world input data. For the example of a typical RGB image of size 256×256, just one layer assigning a single value to each pixel would already require instantiating a weight matrix of size $\mathbb{R}^{3 \cdot 256^2 \times 256^2}$, corresponding to more than 51 GB of memory needed in 32-bit floating point format. To make deep learning for more efficient for image processing, specialised neural network architectures have been developed for specific deep learning applications like computer vision [40].

The most essential architecture used throughout this dissertation is the convolutional neural network (CNN). Based on the concept of convolutional kernels from image processing, they exploit the spatial structure of image data by combining only data from input pixels in a local neighbourhood. Replacing the matrix multiplication in equation (2.1) with a convolution operation yields a *convolutional layer*. For a two-dimensional image $I \in \mathbb{R}^{H \times W}$, the convolution calculates an output feature map F as follows[1]:

$$F_{i,j} = (I * K)_{i,j} = \sum_{i'=-h}^{h} \sum_{j'=-w}^{w} I_{i-i',j-j'} K_{i',j'} \tag{2.4}$$

K denotes a convolutional kernel, indexed by the set $\{-h, \dots, h\} \times \{-w, \dots, w\}$ [40].

For multi-channel input imagery, a separate kernel is applied to each input channel and the results are added together. Similarly, multiple convolution operations can performed for obtaining multiple feature maps. Convolutional layers can be defined analogously for one-dimensional inputs like time series or three-dimensional inputs like voxel data [40].

Convolutional layers are an essential deep learning tool in computer vision due to several desirable properties. First, they require substantially fewer computational resources to store and compute as they only evaluate local connections. Secondly, while each CNN is equivalent to a highly sparse MLP, they tend to be more data-efficient in training, as weights are shared for all locations in the image. So, if a layer was trained

[1] In practice, many deep learning libraries compute a slightly different operation, namely *cross-correlation*. This operation switches the input indexing term to $I_{i+i',j+j'}$. Still, the term "convolution" is used prevalently for both operations in the deep learning context [40].

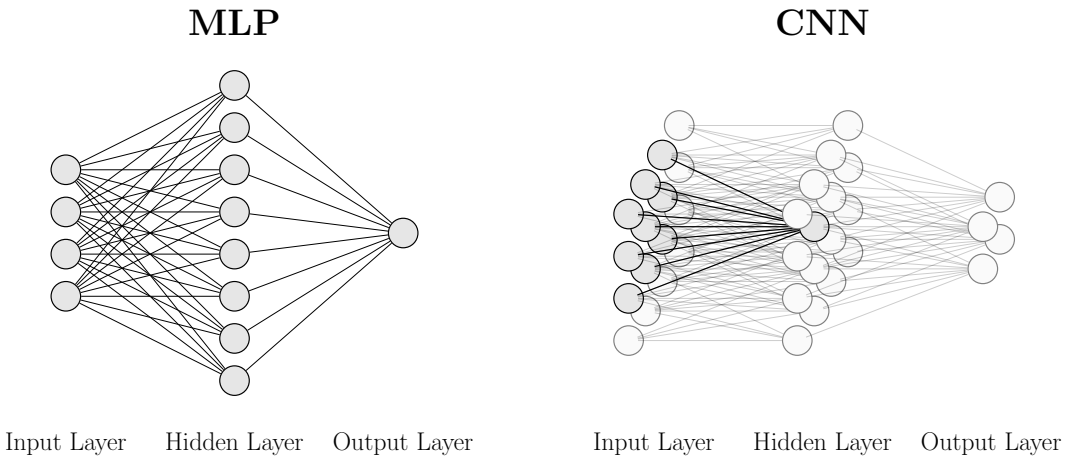


Figure 2.1: Conceptual Architecture of an MLP (left) and a CNN (right).

to recognise a specific feature in one image location, this knowledge would transfer to other locations by design. This *translation equivariance* is a so-called *inductive bias*, which means that the space of possible functions is deliberately limited in order to favour models that adhere to properties believed to be beneficial for the task [46]. As deep learning for remote sensing often deals with image data, CNNs are an indispensable tool for designing neural networks that can solve remote sensing tasks [47].

By passing imagery through a series of convolutional layers, feature maps of increasing abstraction are obtained. While early layers may respond to local features like edges or corners, later layers can detect more extensive features [40]. In order to guide the flow of information through a CNN, two additional techniques are commonly used. The first one, called *pooling*, reduces the spatial resolution of the feature maps. In doing so, the network can connect features from farther apart in the original image without employing huge convolutional kernels. The most common pooling technique is called *max-pooling*. It works by dividing an input feature map into local cells, e.g. of size 2×2, and then reducing each cell to its maximum value [40]. *Skip connections* are another vital design technique for CNNs. They split a network into multiple branches, which are then merged back together through element-wise addition or vector concatenation. This procedure is used in nearly all modern deep learning architectures, leading to more stable training and faster convergence [48].

2.2 Deep Learning in Remote Sensing

Deep learning in remote sensing is a rapidly advancing topic. Conceptually, remote sensing tasks are similar to computer vision tasks. The primary data modality in both fields is rasterised imagery, meaning that the data consists of pixels ordered in a regular grid. Due to this close similarity, the recent breakthroughs in computer

vision based on CNN architectures have given rise to significant leaps in remote sensing methodology [47].

While the fundamental CNN building blocks can be applied to remote sensing imagery, some challenges arise from the large variety of remote sensing imaging modalities. The natural images studied in computer vision generally have three colour channels: red, green and blue. On the other hand, the number of channels varies widely in remote sensing. Multi-spectral images usually have around a dozen spectral bands. Even more extreme, hyper-spectral images can cover hundreds of spectral bands [47].

Many recent CNN architectures employ so-called pre-trained backbones, which means that a large part of the neural network is initialised with parameters that were obtained by pre-training on a large dataset, such as ImageNet [49]. Pre-training aids generalisation and reduces the time needed to train such a model [50]. However, remote sensing imagery is so diverse, with varying channel numbers and resolutions, that pre-trained backbones for remote sensing tasks are often not an option. Instead, researchers tend to initialise the models randomly [47]. So, computer vision tends towards more complex models, relying on the quality of pre-trained backbone features. However, less complex models might be more data efficient, making them better suited for remote sensing tasks.

Computer vision applications in remote sensing are too many to list exhaustively in this chapter. Still, many of them share similar approaches in their methodology and can be reduced to one of four fundamental computer vision techniques: image classification, semantic segmentation, instance segmentation and object detection. Figure 2.2 gives an impression of how each one of these techniques might analyse a given remote sensing image.

Image Classification

The first task where CNNs revealed their potential is *image classification*, where the model must assign one out of several predefined classes to each input image [51]. Computer vision research has explored this task extensively. Datasets like ImageNet [49] serve as standardised benchmarks. New models are often first evaluated on image classification before being evaluated for other tasks [48], [52]. In remote sensing, this task is usually called *scene classification*. Models trained on this task can automatically distinguish between general land use classes [53] or detect specific objects, such as airports or sports stadiums [54].

Semantic Segmentation

Frequently, the location of specific objects within an image is important. Adding a spatial component to the idea of image classification yields the task of *semantic segmentation*. Instead of assigning one label to the entire image, the model must label each individual pixel [55]. One standard benchmark for semantic segmentation in computer vision is the CityScapes dataset [56]. In this benchmark, models must segment objects like cars and pedestrians in urban street scenes. The idea of pixel-wise analysis aligns

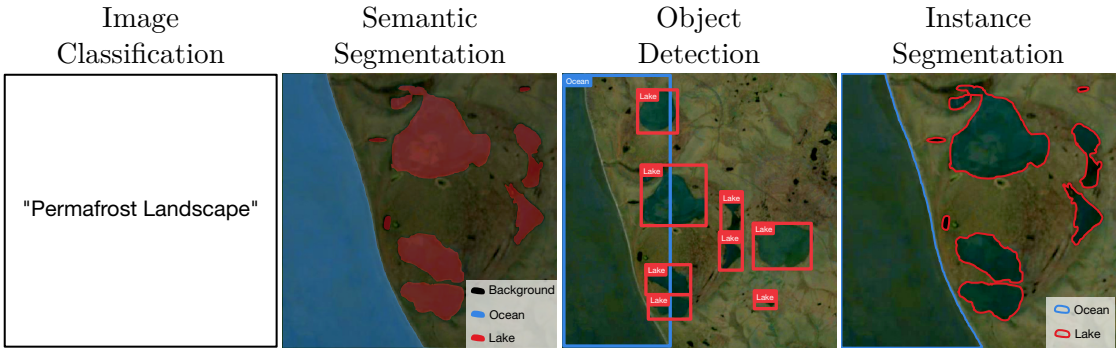


Figure 2.2: Overview of the four main deep learning tasks in computer vision. Given an input Sentinel-2 image from the Bykovsky Peninsula, Siberia, possible outputs for each task setting are shown.

well with mapping tasks in remote sensing, where the location and extent of specific features are needed. Therefore, it is widely applied for various tasks in remote sensing, such as fine-grained land use classification [57] or cloud detection [58].

Object Detection

While semantic segmentation models can be helpful for localisation tasks, they can fail in some scenarios. For example, when multiple objects from the same class touch each other, a semantic segmentation model cannot separate them. In such settings, *object detection* models can be helpful [59]. Here, the predictions are bounding boxes for objects of interest. While such models do not trace the exact boundaries for each object, they can separate individual objects. Therefore, they are helpful when only the number of objects or their rough extent is needed. A standard computer vision benchmark for object detection is the Microsoft Common Objects in Context (COCO) dataset [60]. In remote sensing, object detection models can detect various classes of objects like trees, landslides, ships, or aeroplanes [61].

Instance Segmentation

Some tasks require both pixel-wise mapping and the separation of individual objects. Intuitively, models can conceptually combine semantic segmentation with object detection to achieve this. *Instance Segmentation* trains models to predict a separate mask for each object instance. Following this idea, these models do not only assign class labels to each pixel but also group them into objects ("instances") and separate them from the background [62]. The COCO dataset mentioned above additionally contains instance-level annotations, making it a valuable benchmark dataset for this task as well [60]. In remote sensing, instance segmentation models have successfully been trained to analyse buildings [63] or detect vehicles [64].

These techniques are widely applied in remote sensing, yet they require modifications depending on the exact task to be solved, the satellite platform used, and other factors.

One of the central themes discussed in this thesis will be deciding whether a given task fits into one of the categories above or whether there is a more natural way of formalising the task.

> **Research Question**
>
> Are any of these standardised computer vision approaches well-suited for polar remote sensing tasks, or are there better ways to encode the tasks?

2.3 Remote Sensing in Polar Regions

The polar regions are near the Earth's poles and exhibit extreme climatic conditions. Phenomena like polar night, cloud cover or snow cover have substantial implications on the acquisitions made by Earth observation satellites [38]. Therefore, one must carefully consider which data source to use for polar remote sensing. Multi-spectral optical imaging and synthetic aperture radar (SAR) are the most widely used imaging modes [38]. Similar to regular camera images, multi-spectral images are acquired using optical methods. However, multi-spectral imagery includes more spectral bands, usually in the ultra-violet or infrared ranges [65]. SAR works by transmitting coherent radar pulses to the Earth surface and analysing the radar echo [65]. This section will cover the advantages and disadvantages of these fundamentally different image acquisition techniques for polar remote sensing. In particular, it will also discuss some effects specific to polar environments, as shown in figure 2.3.

Multi-Spectral Remote Sensing in Polar Regions

Optical imaging satellites work based on principles similar to digital cameras. While invisible to the human eye, the spectral ranges beyond visible light contain important information about environmental factors. Therefore, these satellites collect ultraviolet and infrared spectral ranges in addition to the typical red, green, blue (RGB) bands. For these wavelengths, conventional optics and imaging sensors like complementary metal oxide semiconductor (CMOS) sensors are mostly used [65].

Multi-spectral remote sensing is a major branch of remote sensing, with open data programmes by major space agencies providing multi-spectral imagery from their satellites free of charge. The NASA Landsat mission is an example of a long-standing satellite program providing global multi-spectral imagery products. The first Landsat satellite was launched in 1972. Follow-up missions were launched regularly, resulting in a continuous archive of Landsat imagery up until the present [65]. The most recent Landsat satellite, Landsat 9, was launched in 2021 and acquires imagery across 11 spectral bands at varying resolutions between 15 m and 100 m [67]. More recently, ESA's Sentinel-2 mission was launched in 2015 to provide global imagery across 13 spectral bands at resolutions from 10 m to 60 m with a 5-day revisit time [68]. It allows for more fine-grained analysis than Landsat imagery, but covers a shorter time span in turn. Finally, with cheaper launch costs and smaller optics and electronics, private

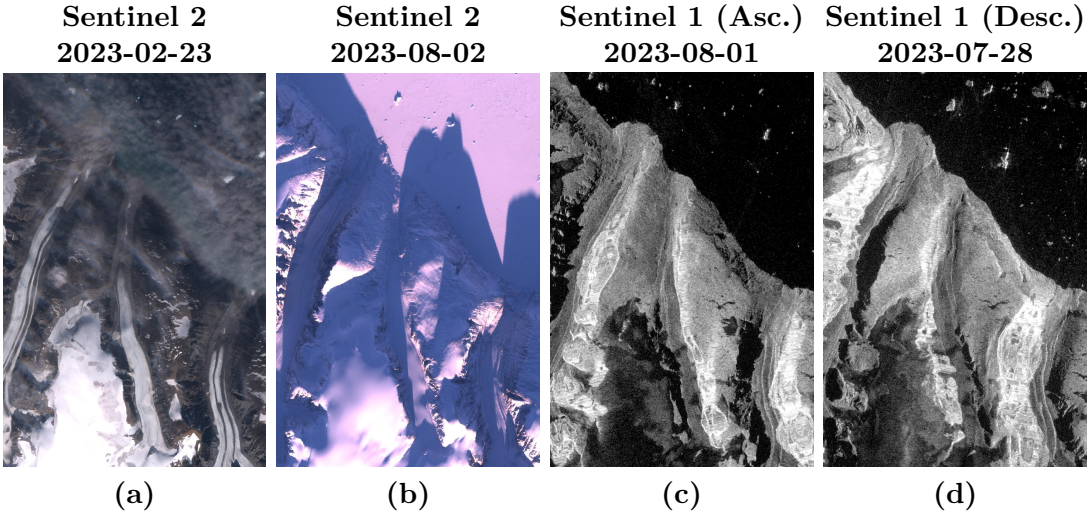


Figure 2.3: Challenges in image acquisition for polar remote sensing. Shown are glaciers in North-East Greenland (Nuussuaq Peninsula). Optical remote sensing can be obstructed by cloud cover (a) or by shadows due to low sun elevation angles (b). Meanwhile, SAR imagery can have missing data in steep terrains due to radar shadow (c-d).

companies have also started operating optical remote sensing satellites, such as Maxar Technologies [69] or Planet Labs [70].

An advantage of multi-spectral imagery is its good interpretability to the human eye, especially for the visual RGB bands. At the same time, the considerable number of spectral channels allows for a differentiated analysis of the objects on the ground, as different materials tend to have characteristic spectral responses [65].

The most prominent drawback of using optical satellite data is the presence of clouds in the imagery. Depending on the region of interest, a significant fraction of all acquisitions can be covered by clouds, effectively preventing any updates during periods with cloud cover [38]. As shown in figure 2.4, this issue is particularly prevalent in some polar regions. Clouds cover over half of the acquired images in regions such as Siberia, East Antarctica, or northern Canada. Figure 2.3 (a) shows how clouds can obstruct glaciers from view in a Sentinel-2 image.

Another drawback of optical imaging is its need for sufficient illumination from the sun. Illumination is not a problem for most regions, as the satellites acquire imagery when the sun's elevation angle is high. For example, the Sentinel-2 satellites acquire their imagery at a local solar time of 10:30 AM [68]. In high latitudes, however, illumination from the sun is not always sufficient. For considerable parts of the year, the sun will not rise in regions beyond the polar circles, a phenomenon called *polar night* [71]. During polar night, optical remote sensing is impossible in these regions, as the sun will not illuminate the ground at all. Figure 2.5 maps the number of days affected by polar night for the Sentinel-2 satellite for the polar regions.

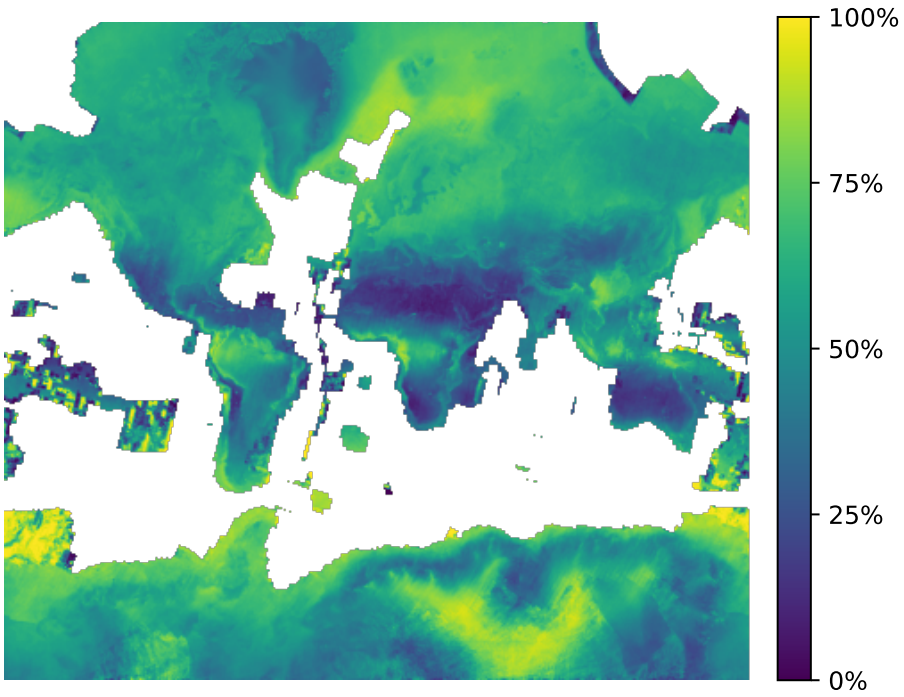


Figure 2.4: Map of Sentinel-2 average cloud probabilities, calculated as the mean of all available tiles from 2015-06-23 to 2023-08-17 in the `S2_CLOUD_PROBABILITY` layer on the Google Earth Engine [22]. Of note are the relatively high values over the continental Arctic in Canada, Siberia, and large parts of the Antarctic coastline.

Even during the polar day, the sun can be at shallow elevation angles for the acquisitions. In such conditions, large parts of the imagery can be covered by shadows, especially in mountainous areas like the Greenlandic fjords. This insufficient lighting and the resulting shadows can lead to poor contrast in essential areas and confuse machine learning models near the boundaries of the shadows. Figure 2.3 (b) shows an example where long shadows from a glacier's fjord boundaries cover most of the glacier area.

SAR Remote Sensing in Polar Regions

Other than optical remote sensing, SAR sensors acquire measurements at much longer wavelengths of the electromagnetic spectrum. For example, ESA's Sentinel-1 satellite works with C-band radar waves at 5.405 GHz, corresponding to a wavelength of roughly 5.55 cm [73]. Radar waves cannot be focussed and sensed using conventional optics. Instead, they are transmitted and received through a radar antenna. A narrow beamwidth is needed when transmitting the signal for acquiring high-resolution data. The beamwidth is inversely proportional to the antenna length. For a finer resolution, the antenna thus needs to be made longer. For the Sentinel-1 satellite, the antenna would have to be multiple kilometres long to achieve the desired 10 m resolution, which

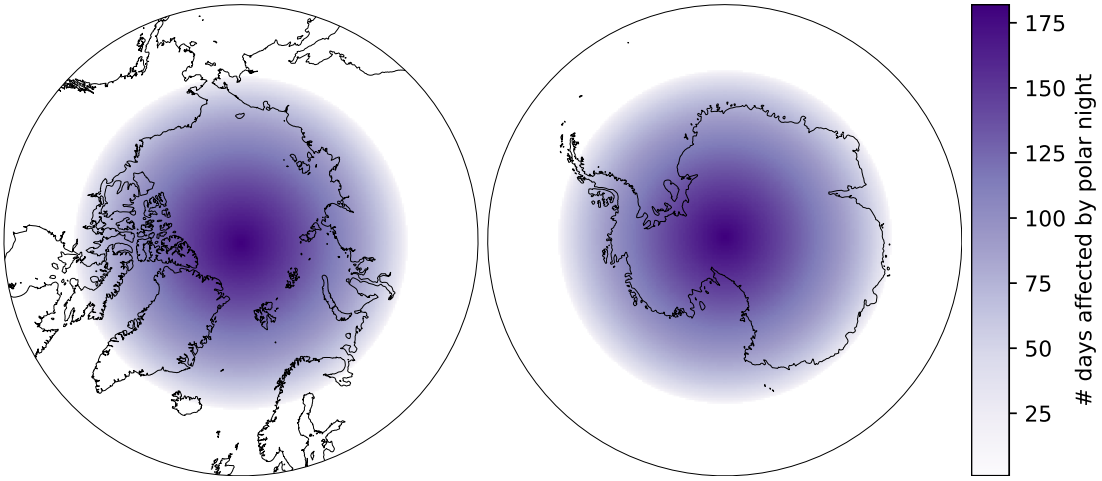


Figure 2.5: Number of days per year affected by Polar Night for Sentinel-2 acquisitions overlaid on maps of the poles. Plot derived using formulae from Meeus [66].

is impractical for a satellite. Instead, the SAR technique combines multiple acquisitions along the satellite's flight path to virtually increase the antenna length [65].

The main advantage of SAR is its ability to penetrate clouds due to its relatively long wavelengths. Further, SAR is an active sensing method, meaning it does not rely on illumination from an external source like the sun. Therefore, SAR can also operate under heavy cloud cover or polar night conditions [38].

SAR sensors acquire measurements in range-doppler space. Due to symmetry, a nadir-looking SAR satellite would receive double echoes from points to the left and right of the ground track, causing an overlay of two areas in the sensed imagery. To avoid this issue, SAR satellites operate with a side-looking acquisition geometry [65]. For example, the Sentinel-1 satellite senses at incidence angles between 20° and 46° [73]. While this side-looking geometry is not an issue for many use cases, it can cause missing data points in the form of radar shadows. These occur in steep terrains when the incidence angle is so high that some areas become occluded by steep surfaces like mountains. so the radar signals cannot reach them [65]. Examples of this are shown in figure 2.3 (c-d), where radar shadows can be observed on both sides of the fjord.

Further, the side-looking acquisition mode of SAR sensors means that a most satellite instruments will only ever look to their left or their right. For most areas on Earth, this does not matter, as the area will be imaged from both directions by acquisitions in ascending and descending orbits. However, side-looking satellites in a near-polar orbit will miss one of the Earth's poles entirely as they always look away from that pole. Left-looking sensors miss the North Pole, while right-looking sensors miss the South Pole [74]. Figure 2.6 shows this issue for the Sentinel-1 satellite, which has a considerable acquisition gap over the South Pole.

As SAR uses polarised radar signals, the type of polarisation can drastically alter the imaging result. Usually, one differentiates this into vertical (perpendicular to the surface) and horizontal (parallel to the surface) polarisation modes [65]. The Sentinel-

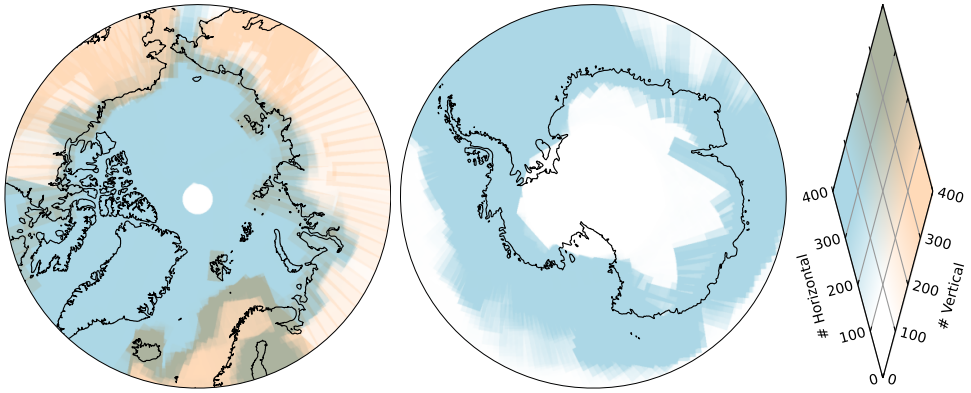


Figure 2.6: Number and polarisation modes of all available Sentinel-1 acquisitions for the polar regions on Microsoft Planetary Computer [72] as of 2023-11-15. Vertically and horizontally polarised acquisition modes are both used for parts of the continental Arctic, The observation gap due to the satellite's right-looking acquisition geometry is visible near the south pole.

1 satellites can operate both in vertical and horizontal polarisations. They generally measure land surfaces in vertical polarisation in the mid-latitudes. For oceans and ice caps, the horizontal polarisation mode is preferred. However, the polarisation mode used for the continental Arctic is ambiguous, with both modes being used at different times [75]. As figure 2.6 shows, horizontal and vertical polarisations are mixed in important permafrost areas like the Siberian coastline and Canada. This can be an issue for machine learning models, as models trained on one polarisation mode do not transfer well to a different polarisation.

3 State of the Art

This section will discuss pre-existing work and parallel developments in the two main tasks addressed in this dissertation, namely glacier calving front detection and permafrost disturbance mapping.

3.1 Calving Front Detection

Glacier calving fronts are a major indicator for the state of the polar ice sheets. These calving fronts are present in marine-terminating glaciers, which are glaciers draining into the ocean. A glacier's calving front is defined as the boundary between ice still attached to the glacier and the ocean or freely floating sea ice [76]. The name calving front derives from the fact that this is the location where new icebergs calve off. Shifts in the calving front hint at underlying melt processes or surge events [77], [78]. Numerous studies have explored remote sensing for mapping calving fronts using various methods, as the calving front is usually well visible in satellite imagery. Many approaches have been evaluated in pursuit of accurate automatic calving front delineations. The following section summarises pre- and post-deep learning methodology for calving front detection.

Traditional Vision Methods for Calving Front Detection

Before the advent of deep learning networks capable of performing pixel-level annotations, calving fronts were primarily detected in satellite imagery using traditional computer vision and statistical methods. Early works started by exploiting the strong contrast between ice and ocean in optical reflectivity and radar backscatter behaviour. For example, Liu and Jezek [79] used this idea to map the calving fronts of the Antarctic ice sheet using SAR imagery. They derived optimal thresholds for each image using a bimodal Gaussian mixture model and then segmented imagery based on these thresholds. Similarly, other unsupervised computer vision methods, such as the watershed algorithm or unsupervised clustering, have been applied for the calving front detection and the closely related task of coastline detection [80], [81]. While these models are generally fast to evaluate, they lack a contextual understanding of glaciers and the ocean. Therefore, features like sea ice, icebergs, or surface melt can confuse the models.

Besides traditional segmentation methods, edge detectors have also been applied for coastline detection and calving front detection. Lee and Jurkevich [82] built a

coastline detection pipeline based on the Roberts cross operator. Similarly, Krieger and Floricioiu [83] use the Canny operator for edge detection and combine it with shortest path finding as a tracing technique in order to find calving fronts in Antarctica. Going beyond convolutional operators into statistical texture analysis, Wang and Liu [84] detect coastlines by analysing local image statistics, which often change abruptly at the boundary between sea and land. Lines of such abrupt changes are then connected using ridge tracing to reconstruct curves that often coincide with the actual coastline.

Other methods, such as active contours or the level set method, can be used to combine assumptions about the ocean and land areas as well as their boundaries, both for coastline detection [85]–[87] and calving front detection [88]. These methods also focus on the boundary between ice and ocean areas, like the edge detectors mentioned above, but also consider properties of the land and ocean areas. Their primary issues lie with stability and robustness related to local minima in the solution space, which are caused, for example, by glacial crevasses. As they are designed to trace a single contour through the imagery, their performance can degrade dramatically if the algorithm takes a wrong turn during the tracing process [1].

Deep Learning for Calving Front Detection

With the introduction of deep learning methods for semantic segmentation tasks, these novel models soon became widely used in remote sensing [47]. CNN segmentation architectures such as SegNet [89] or UNet [55] rely on multi-resolution stacks of feature maps, sometimes called a feature pyramid. These feature maps combine detailed information about individual pixels in the image with more general, high-level information about the overall structure of the image. Due to their solid and robust performance on many tasks, CNNs have become the standard baselines for many mapping tasks in remote sensing [90].

Especially the UNet model has been widely adopted for distinguishing between land and ocean areas [91]. Seeing the strong performance of the UNet model for this task, numerous adaptations add specific layers or modules to the basic UNet structure to improve the performance [92], [93]. The UNet was also readily employed for calving front detection both in Greenland [94], [95] and Antarctica [76]. Tweaking network parameters like the loss function or dropout rate from their default values can improve performance even further [96].

Since deep learning algorithms adaptively learn from annotated examples, they can often arrive at better results than the previously discussed traditional vision methods. Deep learning is the method of choice, especially for increasing robustness against the confounders mentioned earlier, like sea ice or icebergs. However, this comes at the cost of an extensive annotated training dataset. While traditional methods only require tuning a handful of parameters, deep learning models usually have millions of parameters that must be carefully optimised. Therefore, this class of models needs thousands of labelled input images to learn from, which requires a significant up-front time investment from human annotators to generate a training dataset.

Concurrent Research in Calving Front Detection

Parallel to this thesis project, automated calving front detection using deep learning has been an active research area, bringing forth innovative approaches to improve the performance of calving front detectors. Given the strong performance of the UNet baseline models, quite a few works take this model as their starting point. For example, seeing the importance of a broad spatial context, Loebel *et al.* [97] increase the number of down- and up-sampling steps to the UNet. These added layers allow the model to incorporate a larger spatial context into its predictions. Holzmann *et al.* [98] add attention gates to a UNet to arrive at a more interpretable model. They inspect the attention maps and tweak the model hyperparameters to improve performance. Davari *et al.* [99] stack two UNets back-to-back to allow for an intermediate reasoning step. The first UNet predicts the distance of each pixel to the calving front. The second one then predicts the actual calving front from this intermediate output. Similarly, Wu *et al.* [100] interleave two UNets, where one UNet works on the full-resolution image, and a secondary UNet processes a zoomed-out version of the input imagery in order to provide a larger spatial context. The two UNets are linked using attention layers.

As computer vision moved from UNet to more sophisticated segmentation models, the most successful models were also applied in calving front detection. Backbones like Xception [101] enabled broader spatial context windows and extraction of more general features, which has proven helpful for detecting calving fronts [8]. For the segmentation architecture itself, more recent models like DeepLabv3+ [102] can increase the learning capacity of calving front detectors [8], [103], [104].

Of particular interest is the trend towards favouring edge detection over semantic segmentation. While early deep learning models for calving front detection focused exclusively on segmentation models, edge detection has recently become more dominant. It has been used as an additional training objective next to segmentation [8] or as the primary training objective altogether [104]. Incorporating edge detection for calving front delineation is also a central theme of this dissertation and will be discussed in detail in chapter 4. The parallel developments towards edge detection for calving fronts indicate a promising future for this line of research.

3.2 Permafrost Disturbance Mapping

Other than calving front detection, mapping permafrost disturbances in remote sensing imagery with deep learning is not as well explored. In comparison to the glacier movements in Antarctica and Greenland, permafrost changes are less prominent and heterogeneous, seemingly attracting less interest from the remote sensing community. Furthermore, permafrost thaw manifests itself in various disturbances, so the studies for permafrost disturbances tend to be more diverse in the targets they map [39].

Conventional Mapping of Retrogressive Thaw Slumps

retrogressive thaw slumps (RTSs) are the primary permafrost degradation mapped in this dissertation. These features develop on hill slopes and shores in regions of ice-rich permafrost. When the permafrost thaws in these areas, the soil destabilises and starts moving downward [105], [106]. The resulting geomorphological developments fundamentally change permafrost landscapes, as shown in figure 1.2. It is believed that rapid thaw processes like RTSs drive further permafrost thaw and release previously stored organic carbon into the atmosphere [107].

As a prominent manifestation of permafrost thaw, individual RTSs have been studied in field research for many years [105], [108]. Early remote sensing studies used satellite data [109], aerial imagery [110]–[113] or unmanned aerial vehicle (UAV) imagery [114] in order to find the outlines of RTSs for larger regions.

However, manual analysis is infeasible for building a pan-Arctic inventory of RTSs. The first semi-automated studies relied on manually crafted features like the tasseled cap indices [115] to map RTSs, mostly in northwestern Canada [116]–[118].

With machine learning becoming more widely used in remote sensing, these techniques also influenced the research on RTS detection. Traditional classification methods like random forests proved helpful for this task [24]. Bernhard *et al.* [119] combined classifiers like random forests or support vector machines with SAR data to detect RTSs. Nevertheless, the prediction performance of these conventional machine learning models was limited. The large variety of permafrost landscapes and RTS appearances called for more sophisticated analysis methods, such as deep learning.

Deep Learning for Mapping Retrogressive Thaw Slumps

Before this dissertation project, only a few studies tested the feasibility of deep learning for RTS mapping. Huang *et al.* [120] showed strong performance of a DeepLab [50] model for detecting RTSs in the Northeastern Tibetan Plateau. A follow-up study employed the DeepLabv3+ architecture [102] to PlanetScope imagery to achieve similar results for a different region in the Tibetan Plateau [121].

In parallel to this dissertation project, the research on deep learning for mapping thaw slumps has considerably gained traction. Witharana *et al.* [122] applied a UNet model to WorldView-2 satellite imagery in order to map RTS in Banks Island and Ellesmere Island, studying the effects of hyperparameters like input image size. To exploit data fusion, Yang *et al.* [123] merged Sentinel-2 imagery with elevation information and commercial Maxar imagery at a higher resolution. Their study focuses on the Yamal and Gydan peninsulas in Siberia but also includes the training dataset generated as part of the first manuscript in this dissertation [4]. Runge *et al.* [124] use time-series spanning multiple decades to detect RTSs through their temporal dynamics. During the different stages of RTS development, commonly used indices like the Normalized Difference Vegetation Index (NDVI) follow characteristic curves as vegetation first disappears and then slowly regrows over multiple years.

The Tibetan Plateau constitutes a relatively uniform permafrost region regarding its geomorphology. Therefore, it has been the target of several deep learning studies for RTS. Following the feasibility studies mentioned above, some improvements were made to the automated mapping strategies for this region. Huang *et al.* [125] incorporated multi-temporal imagery to map the growth of RTSs over time. Xia *et al.* [126] implemented an iterative training approach. They started with an initial dataset and progressively added new RTS polygons predicted by the network after manual inspection. These new polygons were added to the training set to train a better model. In this way, they mapped a considerably large study area around the Qinghai-Tibet Highway.

A common challenge noted in many works on deep learning for RTS mapping is spatial generalisation. While the studies were able to train deep learning models with solid performance in selected regions, the goal of accurate predictions throughout the Arctic remained elusive. To improve spatial generalisation, Huang *et al.* [127] employed a generative method called *CycleGAN* to generate additional training data. They showed that including this synthetic data in the training process improves spatial generation compared to a model trained only on the original training data.

A first attempt at pan-Arctic RTS mapping was done by Huang *et al.* [128]. Their main data source is the ArcticDEM elevation model [129], which they use to identify changes in elevation and possible headwall lines. A YOLOv4 object detector [130] then uses this data to predict possible RTS bounding boxes. Finally, they validated these boxes through crowdsourcing validation in an online portal.

Mapping of Other Permafrost Disturbances

Besides the aforementioned RTSs, other permafrost disturbances related to thaw have also been explored through remote sensing and deep learning. These studies face similar challenges as RTS detection, most notably the large variety of permafrost landscapes. Therefore, the following paragraphs briefly introduce these other applications.

One direct indicator of thawing permafrost in ice-rich regions are *thermokarst lakes*. The lake basins are formed through melt-induced subsidence and fed by ground ice turning into liquid water [131]. Due to the distinct spectral signature of water compared to other permafrost landforms, they can easily be mapped in remote sensing imagery [24], [132]. These lakes can grow or completely drain over time, giving important hints about sub-surface permafrost processes such as melting ground ice [131].

Ice-Wedges are a unique landform occurring in ice-rich permafrost areas. These wedges form from ice accumulating in soil cracks, leaving the landscape in a polygonal pattern [133]. In a healthy ice-wedge landscape, surface water will accumulate on the boundaries of these polygons. When the ice wedges melt, however, the centres of the polygons will subside, causing surface water to accumulate in the polygon centres instead [134]. Therefore, ice-wedge polygons are a strong indicator for overall permafrost health. Abolt *et al.* [135] proposed the use of CNNs to detect these polygons. In order to analyse the spatial structure of ice wedges, Rettelbach *et al.* [134] extract ice-wedge polygon structures using a graph-based approach. Similarly to RTSs, mapping ice-wedges requires analysing various landscapes across the Arctic.

While the previously mentioned disturbances are primarily symptoms of permafrost degradation, *Arctic Wildfires* are one of the drivers. Even though active fires are only briefly visible in remote sensing imagery, their burn scars and the resulting landscape changes remain visible for multiple years [24]. Nitze *et al.* [24] detected areas affected by such wildfires using satellite image time series from the Landsat mission. Going to higher resolutions, Gibson *et al.* [136] used WorldView-2 data to differentiate between unburned and burned peatlands in western Canada.

4 Exploring Different Representations for Calving Front Detection

This chapter summarises the first part of the dissertation project. In this part, the status quo of approaching calving front detection as a semantic segmentation task is questioned. In doing so, novel ways of encoding calving front detection as different machine learning tasks are found, which lead to better prediction accuracies, higher computational efficiency, and eliminate the need for post-processing steps.

As discussed in section 3.1, early works on deep learning for calving front detection used semantic segmentation models like UNet [55] to segment scenes into land/glacier and ocean pixels. In a second step, the calving front was then extracted by tracing the boundary between these two classes and converting it into a GIS-ready format, like polylines. In experiments with semantic segmentation approaches for calving front detection during the dissertation project, it soon became apparent that this way of phrasing the task is not optimal due to the following reasons:

1. During model training, the loss function emphasises each pixel with the same importance. A neural network model can optimise a considerable fraction of its loss function by correctly classifying regions far away from the calving front, which is very easily done. At this point, the model can then fall into a local minimum of the loss landscape where the simple regions away from the calving front are classified correctly, but the pixels close to the true calving front are highly fluctuating. Such a model will exhibit low loss scores and high pixel-wise accuracies. However, it is rather impractical for extracting the calving front, as the exact location of the front is of high importance in order to detect even slight changes and trends.

2. Considering that the final desired output is a simple boundary line, predicting labels for every pixel in the scene is quite wasteful in terms of computational resources. Also, the post-processing step of extracting the boundary again requires computational resources. Ideally, a model could directly predict the boundary in the desired format instead of taking the intermediate step of predicting a pixel-wise mask. This idea is addressed in section 4.2.

4.1 Learning from Human Annotation Approaches

For the first project of the dissertation, the main goal was to tackle the issue of poor prediction quality near the boundaries. The starting point for this was to closely observe how human annotators go about labelling ground truth data for a given satellite scene. For this, three important observations were made.

1. Human annotators tend to frequently zoom in and out of the image in order to combine large-scale contextual information with accurate local information.
2. When given the option to work with different brush sizes, humans will pick a smaller brush when working in boundary areas, and use a larger brush in uniform regions.
3. Even when asked to draw pixel-wise annotations for `land` and `ocean` classes, human annotators usually start by carefully tracing the boundary between the two classes and then fill in the remaining regions.

Starting with the UNet model as a solid baseline, the approach taken in this first project was to try and mimic the human annotation behaviour in a deep learning model as closely as possible. The result was the HED-UNet model, which constitutes the main contribution of the first included manuscript.

Relevant Publication for this Section

K. Heidler, L. Mou, C. Baumhoer, A. Dietz, and X. X. Zhu, "HED-UNet: Combined segmentation and edge detection for monitoring the antarctic coastline," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022. DOI: 10.1109/TGRS.2021.3064606

Combining Semantic Segmentation and Edge Detection

The first observation is in line with the previous consideration that semantic segmentation might not be the most natural task formulation for calving front detection. In fact, the most important prediction targets are the pixels that lie on the boundary. While not as prominent as semantic segmentation, *edge detection* is another computer vision task that has been approached with deep learning. In edge detection, the goal is to provide a binary output that takes the value zero for pixels on the inside of an object, and the value one for pixels on the boundaries between objects. In a sense, edge detection can be considered as a special case of semantic segmentation with two classes. However, due to the great structural differences between segmenting objects and detecting edges, deep learning models designed for semantic segmentation are not well-suited for this task. Instead, specialized models were proposed for edge detection. One prominent deep learning architecture for this task is the holistically-nested edge detection (HED) model [138]. As a first step towards building a better calving front detection network, the inclusion of edge detection approaches seemed natural. However,

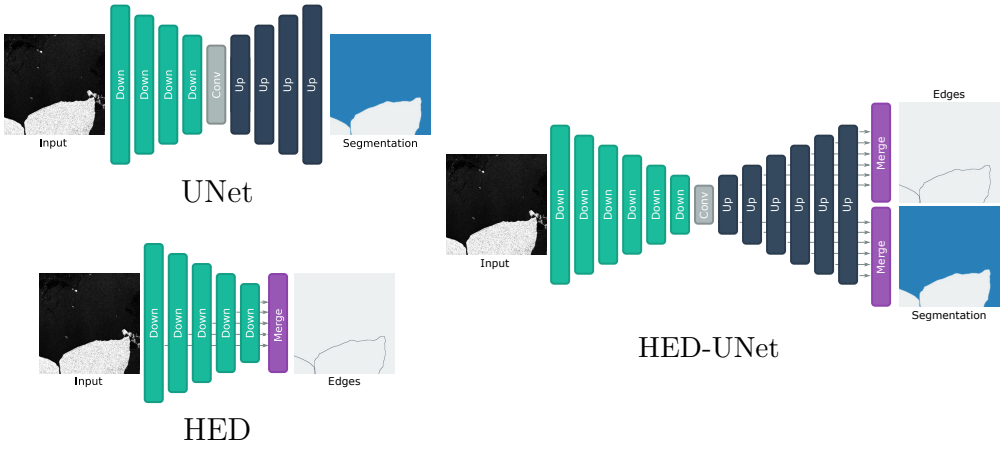


Figure 4.1: Conceptual amalgamation of UNet and HED (left) into HED-UNet (right). While HED merges the outputs of "Down" blocks, HED-UNet uses the outputs of "Up" blocks, allowing for larger receptive fields and deeper decision paths. Figure taken from [137].

when performing only edge detection, important information is lost. For example, it is unclear from an edge detection alone, which side of the predicted boundary is the glacier and which is the ocean. Further, it is possible that the predicted edges become disconnected, posing large challenges for post-processing. In order to combine the advantages of both segmentation and edge detection approaches, a combined framework was built that performs both of these tasks simultaneously.

Initial experiments showed that the previously mentioned UNet and HED models turned out to be promising baseline models for their respective tasks. The developed HED-UNet model therefore is a generalisation of both these models. While the UNet model consists of an encoder and a decoder submodule, HED contains an encoder and a merging head, that combines information from different resolution levels. Combining these two thus results in a model that consists of an encoder, a decoder, and a merging head. To reflect the desired multitask nature, two merging heads were included, one for segmentation and one for edge detection. In order to combine the advantages of both the HED and UNet models, a new network architecture was designed that poses a generalisation of both the UNet and HED models.

The merging head of HED combines data from multiple resolution levels. This pointed towards a way of incorporating the first observation, namely the tendency of human annotators to zoom in and out of a scene. Analogously, merging features from different resolution levels allows the model to combine high-resolution and low-resolution information. While the low-resolution information contains general information like the approximate locations of ocean and land, the high-resolution information is helpful for precise mapping the boundary areas.

When directly adaptating of the HED merging head, the features are merged in a pre-determined fashion. So instead of being able to dynamically mix information as described above, the model can only ever learn fixed coefficients and merge the

features according to these coefficients. But in order to mimic the different brush sizes a human would use, a dynamic merging procedure is be needed. This would then allow the model to select high-resolution information near the boundaries and low-resolution information elsewhere.

The answer to this challenge was found in a completely different research field. At the time of the project, *attention layers* [139] were becoming a popular method in natural language processing. This mechanism allows for a model to exchange information between so-called tokens, which correspond to words or syllables in the context of text. The innovative idea of attention is to dynamically adjust the flow of information in a neural network in a data-driven way, instead of relying on pre-determined weights. Figuratively speaking, the model gains the ability to pay attention to different parts of its input, hence the name *attention layer*. Using the attention mechanism to merge information in the final merging head provided a very natural way of addressing the aforementioned concerns. Instead of paying attention to different words however, the attention merging head in HED-UNet is used to allow the model to pay attention to the different resolution levels of available information. For each location in the output, the attention merging head is thus dynamically aggregating the information from the different feature maps.

Improving Spatial Context

An additional peculiarity of calving front detection in Antarctica is the observation that quite large spatial context windows are needed in some regions to correctly identify the calving front. Confounding features like large icebergs, meltwater or specific types of ice are very similar in local appearance and texture to the opposite class. A study of the receptive field of the baseline model showed that in many instances the model had no chance of correctly detecting the calving front, simply because it lacked enough spatial context.

The largest possible spatial context that a model can take into account for its prediction is called the *theoretical receptive field* of the model [140]. While the theoretical receptive field can be quite large, models still tend to use mostly information close to the output pixel to derive their predictions. An analysis of the resulting *effective receptive field* (ERF) is possible by calculating the gradient magnitudes of the input pixels with respect to a given output pixel, which is usually chosen at the centre of the image. By averaging these gradient magnitudes across the entire test set, it is possible to visualize how large of a spatial context the model is actually using for its predictions [140]. An analysis of these ERFs showed that a baseline UNet model was indeed limited by its small receptive field, as the ERF was sharply cut off by the theoretical maximum. By adding more up- and downsampling layers in HED-UNet as well as incorporating the attention merging scheme, it was possible to greatly widen the spatial context available to the model. Visualisations of HED-UNet's ERF show that the model is indeed making great use of this additional spatial context (A.1, Figure 9).

Experimental Results and Discussion

A dataset of Sentinel-1 imagery of Antarctic calving fronts served as the training and evaluation data for the evaluated models, two regions were reserved for testing, namely Wilkes Land and the Antarctic Peninsula. The PoLiS metric [141] was chosen as the primary metric to measure the average distance between true and predicted calving front line. Interestingly, the baseline results were somewhat divided between these two sites. While segmentation approaches performed better for Wilkes Land (UNet: 271 m, HED: 341 m), an inverted situation was observed for the Antarctic Peninsula (UNet: 483 m, HED: 398 m). As conjectured, the HED-UNet model was able to combine the advantages of both approaches, and showed the best performance for both test sites (Wilkes Land: 222 m, Antarctic Peninsula: 345 m). A more detailed presentation of the results including estimates for the standard deviation across multiple model runs can be found in in the corresponding publication (A.1, Table I).

Through extensive ablation studies, it was further shown that each one of the conducted alterations indeed improved the predictive power of the model. Further, a close inspection of the attention weights revealed that the model indeed learned to perform the conjectured switch between high-resolution features near the boundary and low-resolution features far away from the boundary (A.1, Figure 6).

Deployment and Up-scaling of HED-UNet

As the model proved quite satisfactory for predicting calving fronts in Antarctica, it was decided to implement the model as an integral part of an automated prediction pipeline to derive historical and current calving front positions in Antarctica. The resulting data product, called *IceLines*, was deployed together with the Remote Sensing Data Center at the German Aerospace Center. It contains monthly, quarterly and yearly calving front positions derived from Sentinel-1 imagery for the entire duration of the mission. The dataset descriptor is available at [9], and the data can be inspected and downloaded at `https://geoservice.dlr.de/web/maps/eoc:icelines`.

4.2 Direct Prediction of Contour Lines

While HED-UNet can alleviate some of the issues observed with conventional semantic segmentation models for calving front detection, it still relies on pixel-wise predictions that require intricate post-processing. The second initial question remains: Is it possible to directly output a contour instead of choosing pixel-wise masks as an intermediate representation? If implemented, this change in representation would impose a stronger inductive bias on the model, namely that the calving front should be a consecutive contour. Furter, computational efficiency could be improved, as only the contour would have to be predicted and post-processing could be eliminated.

A literature review of contour prediction methods pointed towards *Active Contours*, or *Snakes* [142] as a promising approach to direct contour prediction. The basic idea behind this method is to start with an initial contour, and iteratively deform it in such

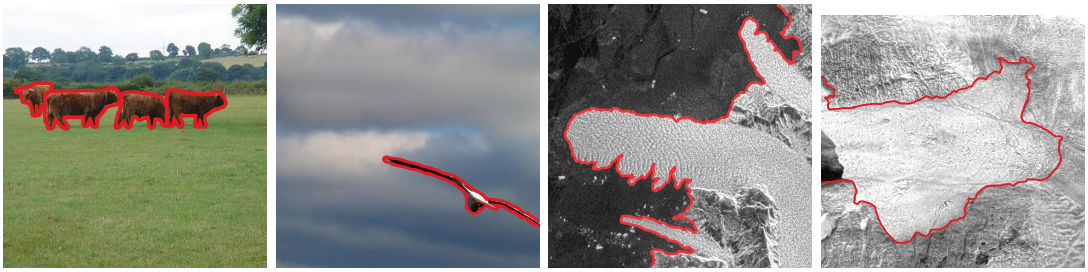


Figure 4.2: Example images from the commonly used PASCAL Visual Object Classes (VOC) dataset [144] (left) for image segmentation, and the CALFIN dataset [8] (right), which was used to train the COBRA model.

a way that it minimizes a pre-defined energy functional, which depends on both the given image and the current contour. This functional is chosen in a way that favours boundary areas in the image, as well as a smooth shape of the predicted boundary. As a conventional computer vision method, this methodology relies on hand-crafted features and a fixed prediction procedure.

Seeing the potential of this approach for a computationally efficient instance segmentation approach, Peng *et al.* [143] adapted this framework for the deep learning age by making it learnable from end to end. Instead of working directly on the raw image values, *Deep Snakes* first employ a two-dimensional CNN in order to extract feature maps of high semantic value. Then, starting with a contour derived from a bounding box, these feature maps are sampled at the locations corresponding to the vertices of the contour, resulting in a sequence of feature vectors. Among these feature vectors, a one-dimensional CNN is then used to pass information between the vertices. Finally, this one-dimensional CNN then predicts offsets to apply to the vertices in order to better match the desired contour.

Seeing the possibility of combining active contours with deep learning, the goal of explicit calving front detection in the form of polylines comes within reach.

Relevant Publication for this Section

K. Heidler, L. Mou, E. Loebel, M. Scheinert, S. Lefèvre, and X. X. Zhu, "A deep active contour model for delineating glacier calving fronts," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–12, 2023. DOI: 10.1109/TGRS.2023.3296539

Design of the COBRA Model

In order to build a calving front detector based on deep active contours, the main ideas behind deep snakes need to be adapted to the setting of calving fronts. Deep snakes were originally proposed as an instance segmentation approach, which means that they are specialized for the detection of rather small, mostly compact objects in an image. In contrast to this, calving fronts often extend beyond the boundaries

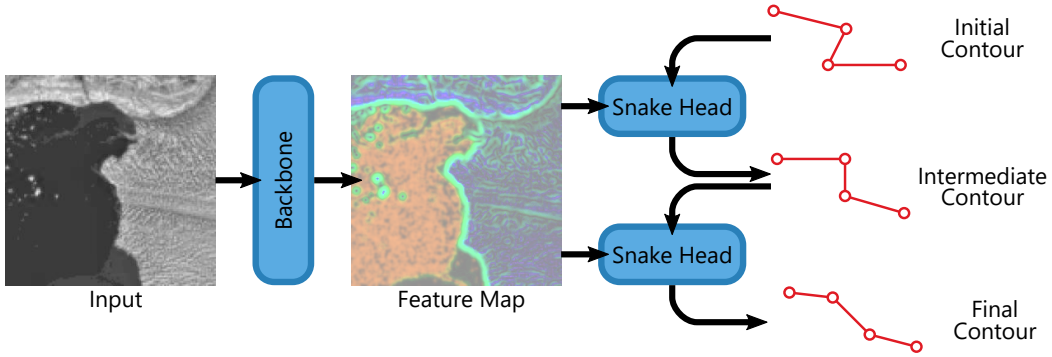


Figure 4.3: Architecture overview of the COBRA model. First, the backbone module extracts feature maps from the imagery. Then, an initial contour is iteratively deformed in a trained manner.

of an image, and there is no clear concept of "inside" or "outside" like in instance segmentation. Further, calving fronts tend to have more jagged, irregular outlines than everyday objects. The conceptual shift between the two tasks is shown by the examples in figure 4.2. This means that the original Deep Snake can not be used directly for the task. Instead, a new model was built inspired by the Deep Snake methodology. The resulting model, Charting Outlines by Recurrent Adaptation (COBRA), implements the concept of iteratively deforming a contour through a neural network. But since the task of calving front detection is quite different from instance segmentation, the model can use a more streamlined architecture.

The deep snakes model was designed to detect multiple objects from various classes in the image, therefore it is constructed as a two-stage model where the first stage comprises an object detector, predicting bounding boxes and classes for the objects in the image. These boxes are then used to initialize the outlines for the snake iteration. This two-stage complexity is not needed in calving front detection. For this task, only a single outline is needed, and there is no need to differentiate between classes of objects. This allows for a greatly simplified pipeline. Since the task definition already states that exactly one contour is needed, the first stage of object detection can be eliminated completely. Instead, the COBRA model directly starts with a single contour that is deformed iteratively. Further, the glacial contours are not closed like in instance segmentation. Therefore, COBRA uses regular one-dimensional convolutional layers in its snake head instead of circular convolutions. Figure 4.3 shows the conceptual overview of the COBRA model. After deriving feature maps using a two-dimensional CNN, the Snake Head iteratively deforms the contour to match the desired output.

Following the observations made during the development of HED-UNet, the COBRA model uses a backbone that can extract features with a broad spatial context. This is important as the model needs to be able to distinguish confounding features like sea ice from actual glacial ice. Instead of the commonly used ResNet [48] backbones, COBRA uses an Xception [101] network with an Atrous Spatial Pyramid Pooling (ASPP) [50]

module as its backbone. Both Xception and ASPP put strong emphasis on deriving features with a broad spatial context. Seeing these advantages, Cheng *et al.* [8] first adapted Xception and ASPP as feature extractors for calving front detection.

Loss Functions for Direct Contour Prediction

Finally, it became apparent that the standard loss functions for regression tasks like the mean squared error (MSE) loss or the L_1 loss have a fundamental issue when used for this task. When discretising a curve as a sequence of vertices, there are many ways to place the vertices along the curve. A common choice is to place the vertices evenly spaced along the curve. Doing so, however, requires a priori knowledge of the total contour length. In contrast to this, the COBRA model is supposed to iteratively develop the contour line, which means that the length of the contour can change throughout the iteration. What is worse, local changes in the contour, like a floating glacier tongue, will change the location of all points on the curve. A visualisation of the global effect of local changes in this setting is shown in figure 4.4. While this problem does not seem to become too apparent in the computer vision application of instance segmentation, it poses a major challenge in calving front detection, partial fractures and crevasses lead to a fractal shape for many calving fronts.

Standard loss functions such as the MSE assume a one-to-one correspondence between the ground truth vertices and the predicted vertices. Figure 4.5 shows why this is impractical for calving fronts. Small reparametrisations like the one introduced in figure 4.4 will shift the vertices, so that the one-to-one correspondence does not line up anymore. This means that the loss function will assign large gradients to vertices even for prediction parts that are matching the true contour, as long as the parametrisations of the curve are not perfectly aligned. The COBRA models initially trained with the MSE loss were therefore punished for not guessing the parametrisation of the curve correctly, even when placing the vertices on the correct contour. This causes the models to predict highly smoothed versions of the true contours, as they are trying to compromise between multiple possible parametrisations of the true contours.

Researching a solution to this challenge, techniques from time-series analysis provided a promising direction. When working with time-series data, many applications require information about the general shape of a time-series, rather than specific time-stamps of peaks or flats in the signal. Therefore, time-series research faced similar issues with re-parametrised signals, and proposed working solutions to compare signals according to their overall shape rather than by exactly overlaying them. Specifically, the technique of dynamic time warping (DTW) [145] is a dissimilarity measure for comparing two time-series. Although it is not a metric in the mathematical sense, it has useful properties for our usecase. DTW considers all possible realignments between the points of the two input time-series by following a set of conditions, namely the realignment being monotonic and surjective in both original time-axes. Among all these realignments, it then chooses the one that minimises the sum of pairwise distances. This sum is then defined as the DTW dissimilarity of the two inputs. When interpreting the vertex sequence of a discretised contour as a time-series, DTW becomes applicable to the con-

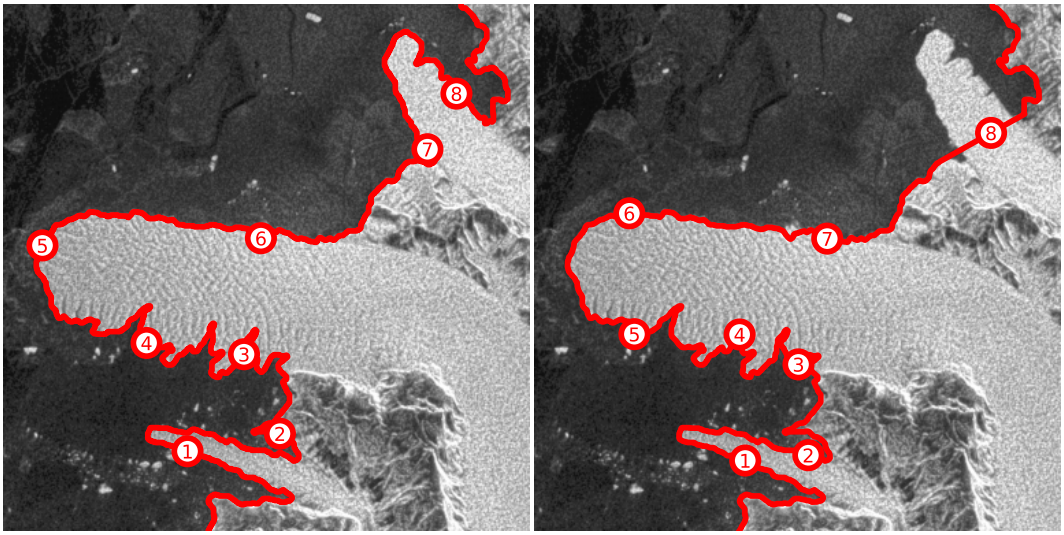


Figure 4.4: Global effect of local changes on the location of discretized vertices for an example from the CALFIN dataset [8]. In (a), the correct ground truth contour is displayed, while in (b), the floating glacier tongue in the top part of the image has been removed from the contour. For both contours, eight evenly spaced vertices are overlaid. The removal of a local structure in the top part shifts all the vertices in the discretized contour.

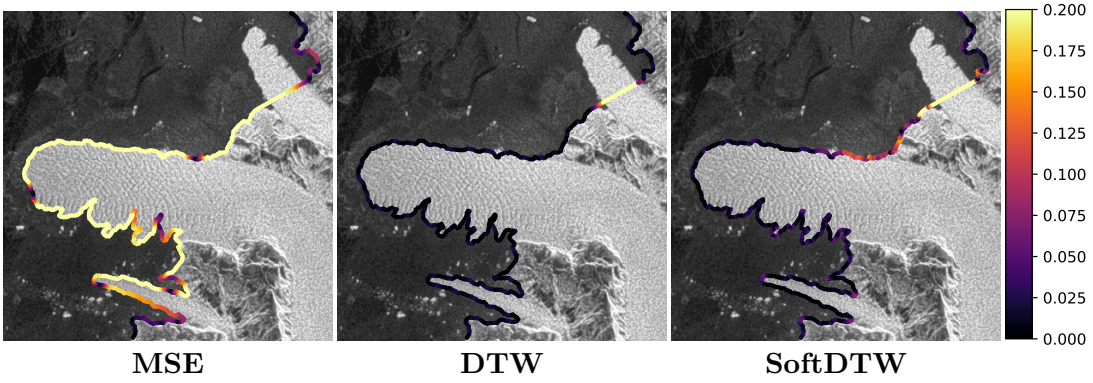


Figure 4.5: Contour gradient magnitudes under different loss functions. Gradients are obtained by comparing the points on the perturbed contour from figure 4.4 (b) to the true points from figure 4.4 (a). While the MSE loss has large gradients nearly everywhere for the locally perturbed contour, DTW and SoftDTW assign very small gradients to the unperturbed contour parts. The regularisation effect of SoftDTW can be observed in regions close to the perturbation.

tour prediction task. The experiments in the corresponding manuscript (appendix A.2, section IV-E) showcase that DTW's advantages are not just theoretical, but translate into more accurate predictions. By replacing the MSE loss with a DTW-based loss, COBRA prediction errors were significantly reduced on all test sets.

However, DTW also has an undesireable property when using it in conjunction with deep learning. Its non-smoothness can interfere with modern deep learning optimisers like Adam [45], which estimate second moments of the loss function. Further, DTW might be too lenient in allowing possible realignments that are very far from the desired equidistant vertex spacing. Facing similar issues, Cuturi and Blondel [146] proposed a smooth generalisation of DTW that replaces the minimum operator with a smooth approximation, namely the softmin operator. This should work better with the aforementioned optimisers and also provide a slight regularisation towards the ground truth parametrisation, as the softmin takes all possible realignments into account instead of just the optimal one. Figure 4.5 shows this regularisation effect for points close to the contour perturbation.

Experimental Results

The COBRA model was trained on the CALFIN dataset [8], which consists of Landsat imagery of the marine-terminating glaciers in Greenland. The model was then evaluated on the CALFIN test set, the TUD dataset [97], and a public subset of the Baumhoer dataset [76]. Once again, PoLiS [141] was used as the main evaluation metric. In order to compare the performance to existing models, pixel-wise detectors like UNet, DeepUNet [97], HED-UNet [1] and the CALFIN model developed specifically for this dataset [8] were considered. Further, adapted versions of the original Deep Snake model [143] and DANCE [147], an improved version of Deep Snake, were evaluated.

In the experiments, COBRA showed quite strong performance, outperforming all other models on the CALFIN and TUD test sets. For the Baumhoer test set, performance was slightly worse than that of the DeepUNet model. This was conjectured to be caused by the higher complexity of the Antarctic calving fronts in the Baumhoer dataset compared to the ones in Greenland. And indeed, an ablation study showed that when doubling the number of contour vertices, COBRA can achieve competitive accuracy also on this dataset.

Uncertainty Estimation

Deep learning models are highly complex with millions or even billions of parameters. Therefore, statistical evaluations of the model uncertainty are not as easily done as they are for simple models like a linear regression. What is worse, deep learning models tend to be overconfident in their predictions, as the ground-truth presented during training is generally definite, which means that the model is never presented with ambiguous cases during training [148]. But especially in Earth science contexts, the reliability of the data is highly important for the downstream use of the model predictions. Therefore, this

study also explores how to quantify the model uncertainties for calving front detection, and whether the COBRA approach can be helpful in this regard.

The output of a segmentation model, a segmentation mask, can be statistically modelled as a collection of individual random variables. In order to estimate properties like the uncertainty of this collection, information about the joint probability distribution for all predicted pixels is needed. However, even for moderate image sizes, this means working in a space with thousands of dimensions. When changing the representation of the calving front from a segmentation mask to an explicitly parametrised contour, the number of predicted variables is drastically reduced. Therefore, one additional hypothesis was pursued in this project, namely the idea that explicit contours might be a better target for uncertainty quantification methods than segmentation approaches.

It is quite challenging to quantify the accuracy of model uncertainty estimates due to the randomness involved. In order to get some indication on the quality of model uncertainty predictions, the following assumptions were made for the evaluation on the test data: Whenever the predicted uncertainty is low, the model prediction should be close to the ground truth. Conversely, if the model prediction deviates far from the ground truth, the predicted uncertainty is also expected to be large. This can be numerically evaluated by computing the Pearson correlation between the predicted model uncertainty and the model error on the test dataset.

For quantifying the uncertainty of the model predictions, the commonly used *Monte Carlo Dropout* [149] method was used. In the experiments, the correlation between predicted uncertainty and model error was indeed highest for the COBRA model on two of the three test datasets, suggesting that the explicit contour representation might indeed be helpful for quantifying uncertainties in its predictions.

Applying COBRA on a Large Scale

Relevant Publication for this Section

T. Li, K. Heidler, L. Mou, Á. Ignéczi, X. X. Zhu, and J. L. Bamber, "A high-resolution calving front data product for marine-terminating glaciers in Svalbard," *Earth System Science Data*, vol. 16, no. 2, pp. 919–939, 2024. DOI: 10.5194/essd-16-919-2024

In a follow-up study, the COBRA model was used to derive a dataset of calving fronts in Svalbard of unprecedented temporary resolution. As COBRA was designed and trained on glaciers in Greenland, it was not clear from the beginning whether the model would generalise well by itself, or whether fine-tuning would be necessary. These worries turned out to be unfounded, however, as the model was predicting calving fronts with high accuracy without any additional changes.

Using COBRA as part of a fully automated pipeline, 149 marine-terminating glaciers were analysed for the time-span from 1985 to 2023. For this, optical data from the Landsat, Terra-ASTER and Sentinel-2 missions was used. Further, Sentinel-1 SAR data was included to allow for observations even in cloudy conditions. During this study, the

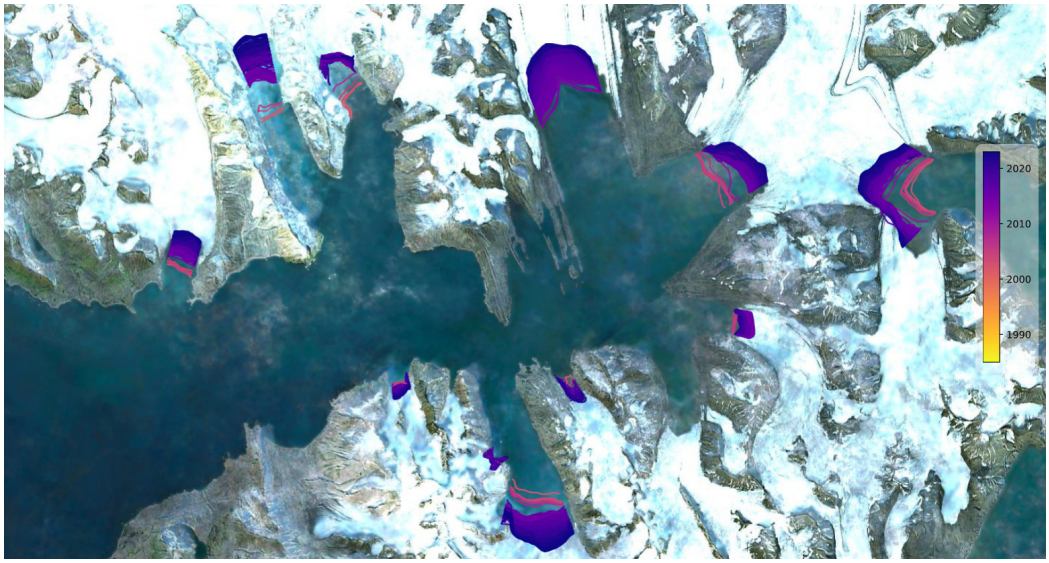


Figure 4.6: Examples for calving front lines from the Svalbard study. Traces for 11 glaciers in Southern Svalbard are shown for the time span from 1985 to 2023. Mostly, retreating glacier developments can be seen. Background satellite image: ESA Sentinel-2. Interactive version of the map: `https://maps.heidler.info/svalbard`

COBRA model proved flexible for using different input modalities like SAR and optical imagery. COBRA's direct output of contour lines allowed for more efficient processing, as the contours were ready for further analysis without further post-processing. The resulting dataset consists of a total of 124,919 calving front positions.

In order to assess the quality of the derived data product, the calving front positions were compared with an existing dataset. Moholdt *et al.* [150] provide manually digitised calving fronts on an annual basis for the years 2008 to 2022 through the Copernicus Glacier Service. By matching same-day calving fronts from this dataset with our dataset, it was possible to calculate the average distance between calving front traces from the two datasets. With an average mean distance error of 32 m, our data product aligns rather well with the data provided by Moholdt *et al.* [150]. A comparison of the calving front change rates between the two data products also showed a very good agreement at an R^2-score of 0.98.

One central measure of interest was the overall advance or retreat of individual glaciers. To summarise the glacial movements into a single number, the derived contours were combined with glacier metadata such as the central flow line. In this way, the movement of the calving front along this central flow line could be measured as a single, interpretable number. In this way, 123 of the analysed 149 glaciers were identified as showing retreating behaviour, while 16 showed an advancing trend, excluding surging glaciers. The remaining 10 glaciers showed surging behaviour. During surging events, these glaciers are rapidly speeding up their flow for a short amount of time. These surges are not caused by external climatic factors, but instead by internal conditions

within the glacier [78]. The high temporal frequency of the observations generated by this study allows for the automated detection of such surge events, which might eventually lead to a better understanding of the underlying processes.

Relating the observations from this dataset of calving front positions with environmental factors like ocean or air temperatures should allow for a better understanding of the behaviour of the glaciers. The unprecedented sub-seasonal resolution allows for assessing not only long-term trends, but also seasonal variations in glacier dynamics.

5 Learning to Map Permafrost Disturbances from Limited Labels

For the second part of the dissertation project, the goal was to accurately map so-called Retrogressive Thaw Slumps (RTS) across the continental Arctic. These permafrost disturbances are mass movements akin to landslides and are indicators for overall permafrost health. Given an input satellite image, a well-performing model should be able to detect all RTSs present within the image. Other than the task of calving front detection, the focus lies indeed on the areas themselves and not the boundary between them. Therefore, the RTS detection task was approached with binary semantic segmentation. The training objective for these models is therefore to predict the output value 0 for non-RTS pixels and the value 1 for RTS pixels.

However, using deep learning for RTS detection comes with its own challenges. Prior to this dissertation project, only limited studies had been conducted for RTS mapping with deep learning, as discussed in section 3.2. These existing studies focused on one geographic region, the Tibetan Plateau. Compared to this specific region, the entire Arctic is far more diverse in terms of landscapes and their appearances. The scientific contributions in this chapter aim to pave the way for models that can predict the presence of RTS anywhere in the Arctic. In order to train a deep learning model with strong performance in pan-Arctic RTS mapping, a couple of challenges need to be overcome.

1. *Classification Ambiguity:* It can be hard to tell from just a satellite image whether a feature is an RTS or not. Even permafrost experts often disagree on specific features, with complete disparity in some regions [151]. In order to confirm their classifications, experts often look at additional information such as time-series data or elevation information. Further, the spatial context is important, as RTS formation usually requires a drain such as a river, a lake, or the ocean.

2. *Label Imbalance:* RTSs only make up a small fraction of the Arctic regions. The dataset built in the first manuscript for this chapter (A.4) includes only areas known to contain a large number of RTS. Still, less than 1% of the overall study area analysed for this dataset is covered by RTS. Most deep learning algorithms were designed with roughly uniform class distributions in mind. Therefore, deep learning can struggle considerably with such class imbalance.

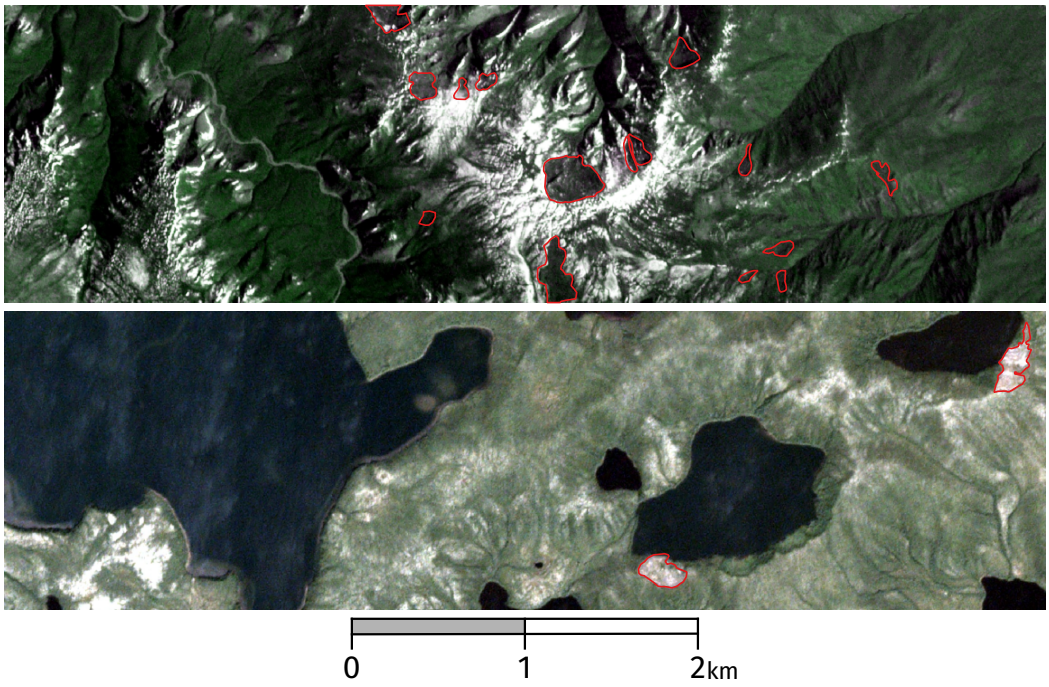


Figure 5.1: RTS mapping examples. Both images show PlanetScope imagery with RTS outlines overlaid in red. Top: Horton site. Bottom: Lena site.

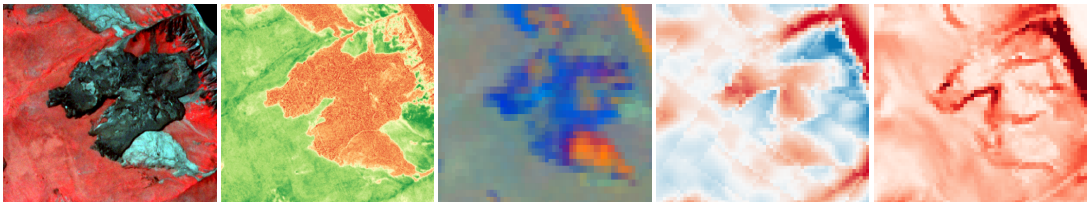
3. *Spatial Generalisation:* The high variability of landscapes across the Arctic poses a major challenge for the spatial generalisation of the models. RTS grow according to similar mechanisms all throughout the continental Arctic. However, RTS features and the surrounding landscapes can look drastically different in various parts of the Arctic. This is due to variations in factors like soil type and vegetation.

5.1 Feasibility of Deep Learning for RTS Mapping

Before developing domain-specific models, a first feasibility study was conducted on deep learning for pan-Arctic RTS mapping. The goal of this study was to gauge the usefulness of deep learning for this task, as well as understanding the effect of different model architectures and training procedures.

Relevant Publication for this Section

I. Nitze, K. Heidler, S. Barth, and G. Grosse, "Developing and testing a deep learning approach for mapping retrogressive thaw slumps," *Remote Sensing*, vol. 13, no. 21, p. 4294, 2021. DOI: 10.3390/rs13214294



Planet Optical (False Colour) | Planet NDVI | Tasseled Cap Trends | Relative Elevation | Slope

Figure 5.2: Overview of data modalities for RTS detection displaying a major thaw slump. For this feature, clear signatures can be seen across all modalities. Visualisation adapted from [152].

Study Sites and Dataset

For this, a dataset of RTS ground-truth polygons was built. Six study areas were chosen for their abundance in RTS with care to represent different parts of the Arctic. These study sites are Banks Island, Herschel, Horton and Tuktoyaktuk in northwestern Canada, as well as the Kolguev site in northwestern Russia and the Lena site in central Siberia. Figure 5.4 shows the locations of these regions within the Arctic.

For each of these sites, multiple PlanetScope satellite scenes were acquired, and the outlines of the contained RTSs digitised manually. Each annotation was checked by another expert in order to ensure high quality of the training labels. Overall, 142 PlanetScope scenes from 2018 and 2019 were analysed and 2172 RTS polygons were mapped for this study.

RTS can be difficult to discern using just optical imagery. Therefore, the input data was enriched by including auxiliary data. First, the NDVI was calculated from the red and near-infrared bands as a measure of vegetation state. Further, as RTS are a dynamic phenomenon, aggregated trend information derived from Landsat timeseries was included as an additional input. These timeseries were aggregated by first computing the tasseled cap indices [115] and then computing their regression slopes over the last 20 years. Finally, as RTSs are hillslope processes, elevation data from the ArcticDEM [129] was also included. In order to prevent overfitting on absolute elevation values, only the slope and relative elevation were used. All input layers were stacked together and the polygon masks were rasterised into a binary mask. These data stacks were then cut into tiles of 256×256 pixels for efficient batching during model training.

For the deep learning experiments, three model architectures were evaluated. The first, UNet [55] is a commonly used baseline model for segmentation tasks in remote sensing data (cf. chapter 4). The second model, UNet++ [153] is a derivative of UNet obtained by coupling the encoder and decoder more tightly through additional convolutional modules. Finally, DeepLabv3 [50] is a more recent semantic segmentation model that incorporates new ideas like atrous convolutions and features a tailored ASPP module for combining information at different scales.

All three architectures can be combined with an arbitrary CNN encoder architecture for feature extraction. Here, three different configurations of the commonly used ResNet [48] were evaluated, namely ResNet-34, ResNet-50 and ResNet-101. Altogether, 9 combinations of encoder and segmentation architectures were evaluated.

One focus point of this study was to evaluate spatial generalisation. Therefore, the splitting of the data into training and test data was done on a regional level in a leave-one-out cross-validation scheme. Sor for each training run, one region was excluded from the training data. After the training process, the model was then evaluated on this excluded region.

Sparsity of RTS

A major challenge in mapping RTS with deep learning is their sparsity. As mentioned earlier, less than 1% of all pixels in the dataset belong to the RTS class. This means that the negative background class dominates the dataset. When training any of the deep learning models without additional changes, the models quickly converge towards predicting all pixels as belonging to the background class. In terms of the model's loss landscapes, constantly predicting the background class appears to be a local minimum that a model can fall into during training. While such a model has more than 99% accuracy, it is not useful for the task at hand. In order to encourage the model to put more focus on the underrepresented positive class, two strategies were employed: A different loss function and a change to the training schedule.

With regards to the loss function, the commonly used *cross-entropy loss* is known to struggle with class imbalances. The alternative *focal loss* [154] addresses this by putting a larger weight on wrongly classified samples and less weight on samples that are already classified correctly. This is achieved by modifying the standard cross-entropy slightly:

$$\mathcal{L}_{\text{Focal}}(p_{\text{true}}) = -(1 - p_{\text{true}})^{\gamma} \log(p_{\text{true}}) \tag{5.1}$$

Here, p_{true} denotes the predicted probability for the true class of a sample and γ is a focusing parameter. For $\gamma = 0$, the regular cross-entropy is recovered. When increasing the value of γ, the described re-weighting behaviour is obtained. In our study, γ was set to 2 as recommended in the original paper [154].

As a second strategy for tackling class imbalance, the training schedule was changed. For the first 100 epochs of training, image tiles without any positive target pixels were excluded from the training. This increases the fraction of positive pixels considerably and therefore encourages the model to start predicting positive pixel labels. However, this procedure favours false positives, as the model does not learn that there can be tiles without any targets. In order to rectify this overrepresentation, the model is then trained for 20 more epochs on the full dataset, in order to decrease the amount of false positives. Experiments show that this two-stage training procedure greatly improves model performance over just training on the full dataset for the same time.

Experiments and Results

Based on the previous considerations, we conducted deep learning experiments on the dataset. Due to the heavy class imbalance, pixel-wise accuracy is a poor indicator for actual model performance. Recalling that less than 1% of the pixels contain positive targets, a model that constantly predicts the background label for all pixels will already reach more than 99% accuracy. Therefore, different metrics are needed to measure prediction quality in such cases. The intersection over union (IoU) for the positive class was chosen as the primary evaluation metric for this study.

For each of the nine model configurations and each of the 6 cross-validation regions, we trained 100 models starting from a random initialisation. Overall, the UNet++ models performed the best. For the backbones, there were no clear trends as to which performed better, leading to the conclusion that the smallest one, ResNet-34, already has sufficient capacity for this task.

For the Horton, Kolguev and Lena regions, the best models were able to reach solid IoU scores of 0.55, 0.48 and 0.58, respectively. For the remainaing regions, results were less encouraging. The best models for the Banks Island and Herschel study areas, the best models reached an IoU of 0.39 for both sites. Finally, the Tuktoyaktuk region proved as the most challenging for all models, so that even the best model only reached an IoU of 0.15. These numerical results underline the diversity of the permafrost regions and the need for better spatial generalisation. More detailed evaluation results can be found in the corresponding manuscript (A.4).

Another observation from this study was the models' lack of training stability. The top-performing models exhibited rather good IoU scores and the visual appearance of the model predictions was convincing. However, model performance was strongly fluctuating in some regions. The random nature of model initialisation and batch sampling during training heavily influenced the performance in these regions.

Both of these observations suggest that the training dataset used in this study is not large enough to thorougly represent the various appearances of RTSs and their surrounding landscapes. The goal for follow-up research was therefore to address this issue in order to train models that train more robustly and generalise better.

5.2 Semi-Supervised RTS Mapping

The permafrost-underlain Arctic spans vast areas, estimated to make up more than 10% of the Earth's land surface [26]. Therefore, annotating a substantial fraction of these areas for training purposes is not a viable option. Instead, the study outlined in this section proposes a new, efficient way to additionally extract information from unlabelled data. In this way, model performance can be improved even with limited labels. The study explores semi-supevised learning for semantic segmentation. This means, that the model is trained on both labelled and unlabelled imagery at the same time. By enforcing some desirable properties on the predictions for unlabelled imagery, the model can learn to generalise better to new regions, as it will have seen more variety in training examples than a model trained just on a small, labelled training dataset.

Relevant Publication for this Section

K. Heidler, I. Nitze, G. Grosse, and X. X. Zhu, "PixelDINO: Semi-supervised semantic segmentation for detecting permafrost disturbances," *IEEE Transactions on Geoscience and Remote Sensing (in review)*, 2024. DOI: 10.48550/arXiv.2401.09271

Semi-Supervised Semantic Segmentation

The term *semi-supervised learning* describes machine learning methods that combine supervised learning with additional unlabelled data [155]. As training data is often a limiting factor for deep learning tasks, various approaches have been proposed for this. These approaches can be broadly grouped into the following three categories:

1. **Consistency regularisation** encourages the model to behave consistently under a certain class of perturbations in addition to training on the labelled data. These perturbations can be done in the input space through data augmentations [156] or by interpolating between samples [157]. Other approaches manipulate the feature space with dropout [155], or by adding noise [158], [159]. Across these perturbations, the model is then regularised to be consistent in its final network outputs [155], [160], [161] or in the feature space [158].

2. **Adversarial semi-supervised learning** uses ideas from generative adversarial networks (GANs) [162] to learn from unlabelled data. One approach for this is training the network to convince a discriminator network that its predictions were are actually ground truth data [163]. Another approach is to use GANs to generate additional synthetic training data [164].

3. **Self-supervised pre-training** can also be regarded as a building block for semi-supervised learning. Instead of randomly initialising the network weights from scratch, the model is first trained on a pretext task on a large, unlabelled dataset. Then, the task-specific dataset is used to fine-tune the network to solve the given task [165]–[167].

The possibility of improving model performance by including unlabelled data aligns well with the lack of labelled training data identified as the main challenge in section 5.1. Applying semi-supervised learning for the semantic segmentation of RTSs is therefore the goal of the study presented in this section. Compared to image classification, semi-supervised learning for semantic segmentation is not as well explored [156], [163], [168].

Initial experiments applying semi-supervised learning techniques to RTS detection showed some promise, but did not lead to satisfactory improvements in performance. A possible explanation for this is the sparsity of RTS targets. Most consistency regularisation methods enforce consistency through the labels. In RTS detection, there are only two classes, of which one dominates the vast majority of the area. Therefore, only

very little information can be passed through enforcing label consistency. Similarly, adversarial approaches do not seem to help much. The large fraction of background pixels impedes the adversarial training setting. A generator can rely on synthesising completely empty background tiles, as the discriminator cannot rule them out as fake.

Pixel-wise Self-Distillation

If it were possible to employ consistency regularisation over a different set of classes than the ones prescribed by the RTS mapping task, the model would receive a much stronger training feedback from the unlabelled images. As no class information is available beyond RTS targets, these additional classes have to be somehow defined by the model itself. We will call such synthetic classes that have no pre-defined semantic meaning *pseudo-classes*.

There are two major challenges in having a model generate a consistent classification scheme. First, a constant assignment will always be consistent. A simple solution for the model is to assign the same class label to all input pixels. Naturally, this assignment will be consistent across any given perturbation of the input data. However, such a model will not learn useful features from the consistency regularisation. Therefore, the first challenge is to ensure *variability* in the model predictions, encouraging the model to make use of all pseudo-classes. But, even when ensuring diversity, there is another undesired mode of consistency that the model can converge towards. This happens when the model constantly outputs a uniform mixture of all available classes. In this way, the model avoids having to decide for a certain class to assign to each pixel. These predictions do in fact make good use of all possible classes, but they are still not informative. The second challenge is therefore to ensure *sharpness* of the model's predictions [169].

Caron *et al.* [169] approached very similar challenges in their work on self-distillation with no labels (DINO). They apply the concept of consistency with pseudo-classes to the self-supervised representation learning for image data. In their training pipeline, they first create two augmented versions of an input image. The first augmented version is then run through a teacher network. Before applying the final softmax operation, the resulting model outputs are centered around the mean and sharpened by re-scaling them. The centering operation prevents individual classes from dominating the predictions, enforcing variability among the pseudo-classes. Meanwhile, the sharpening amplifies any tendencies away from a uniform distribution for the predictions. Taking this teacher output as a training label, the student model is then trained to predict the same output on the second augmented version of the input image. Finally, the teacher's weights are kept as an exponential moving average (EMA) of the student's weights. In this way, models can learn image features of high semantic value without having to rely on labelled data. Downstream experiments on benchmark datasets show that models pre-trained with DINO are competitive with models pre-trained on large annotated datasets [169].

The DINO framework learns to assign a single pseudo-class for the entire input image. For semantic segmentation, however, the model needs to also understand precise

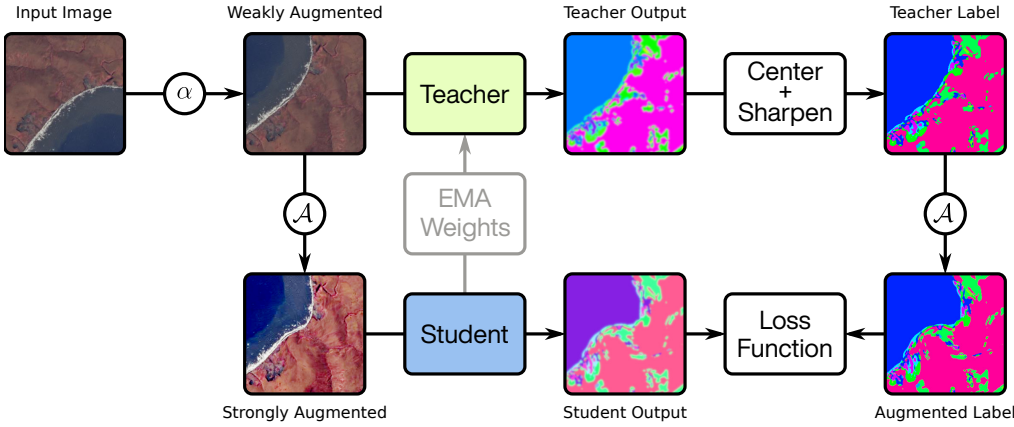


Figure 5.3: Workflow for PixelDINO to learn from unlabelled imagery. An unlabelled input image is weakly augmented. The teacher network then predicts pseudo-classes for each pixel in this image. After centering and sharpening the teacher's predictions, these pseudo-classes are stronly augmented together with the weakly augmented image. This pair is then used as a labelled sample for the student to learn from. The teacher's weights are continually updated as the EMA of the student's weights. Figure taken from [5].

information about individual locations within the image. The proposed PixelDINO framework therefore takes the idea of self-distillation with no labels to the pixel-wise level. Instead of predicting a pseudo-class for the entire image, teacher and student are now replaced by semantic segmentation models, and predict a pseudo-class for each individual pixel.

When adapting the DINO concept to semantic segmentation, one major challenge arises. In the classification setting, data augmentations do not change the label. For semantic segmentation, however, geometric transformations like rotations will change the locations of objects in the image. Therefore, the segmentation masks need to be augmented by the same geometric transformations as well. In the original DINO setting, two independent augmentations are used. Therefore, it is not possible to transfer the segmentation maps from the teacher to the student. Instead, we make use of an idea proposed by Upretee and Khanal [156], namely to employ two consecutive augmentations instead of having them independent from each other. The intermediate version after the first augmentation is the version used by the teacher. The version obtained after having applied both augmentations is then the one used by the student.

The inputs to teacher and student network serve different purposes. In order to ensure high quality of the pseudo-class labels generated by the teacher, the teacher's input imagery should be easy to analyse without too many distortions. Meanwhile, the student network should learn to segment even strongly distorted inputs. To account for these different requirements, the concepts of *weak* and *strong augmentations* were introduced [160]. Weak augmentations, denoted by α, are data augmentations that

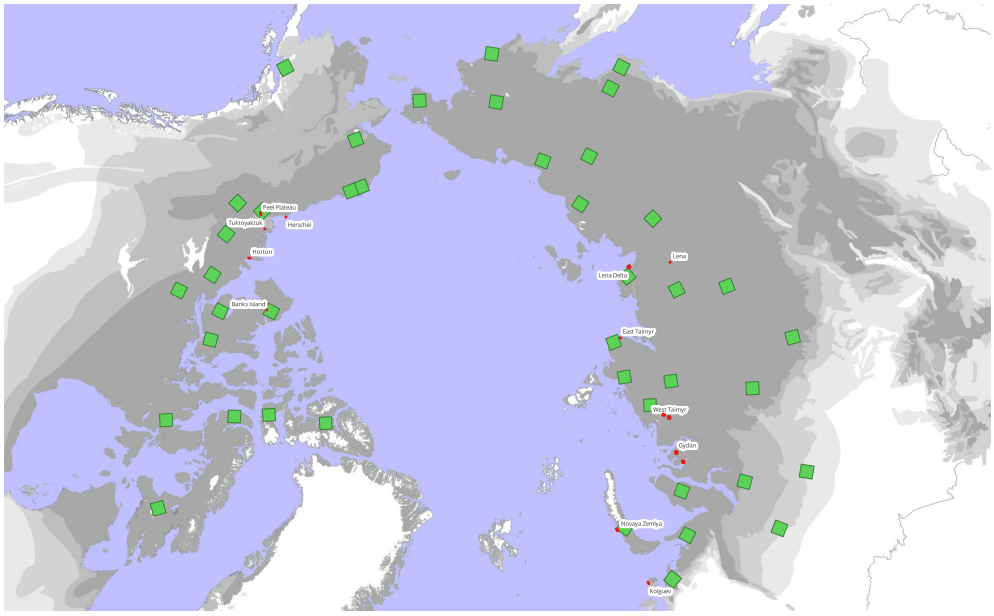


Figure 5.4: Distribution of training sites used for RTS mapping. Labelled sites (red) cover only a small fraction of all permafrost regions (gray), leading to poor spatial generalisation. By including unlabelled sites (green) into the training, as explained in section 5.2, model performance is greatly improved. Figure taken from [5].

do not increase the difficulty of analysing a given image. This class of augmentations includes geometric operations like mirroring the image or rotating it. Strong augmentations, denoted by $\mathcal{A}$, are then any augmentations that distort an image in such a way that makes it more difficult to analyse. This includes operations like blurring an image, changing the colour-space by adjusting brightness, hue or contrast, cropping and resizing of the image, as well as distorting the image geometry through warping operations. The PixelDINO training procedure using weak and strong augmentations is shown in figure 5.3.

In order to train a model in a semi-supervised setting, PixelDINO training is combined with regular supervised training. During each model training step, both a batch of labelled imagery and a batch of unlabelled imagery are taken. For the labelled imagery, a supervised training step is performed, while for the unlabelled imagery, the PixelDINO procedure is performed as the training step. The resulting model is therefore trained on both objectives in parallel, which is expected to provide stronger results than supervised training by itself. An algorithmic description of the full semi-supervised PixelDINO training procedure can be found in the corresponding manuscript (A.5, Algorithm 1).

Experiments and Results

By the time of this second study, the dataset from section 5.1 had grown by a few more regions. In Canada, the Peel Plateau was added as an additional study site.

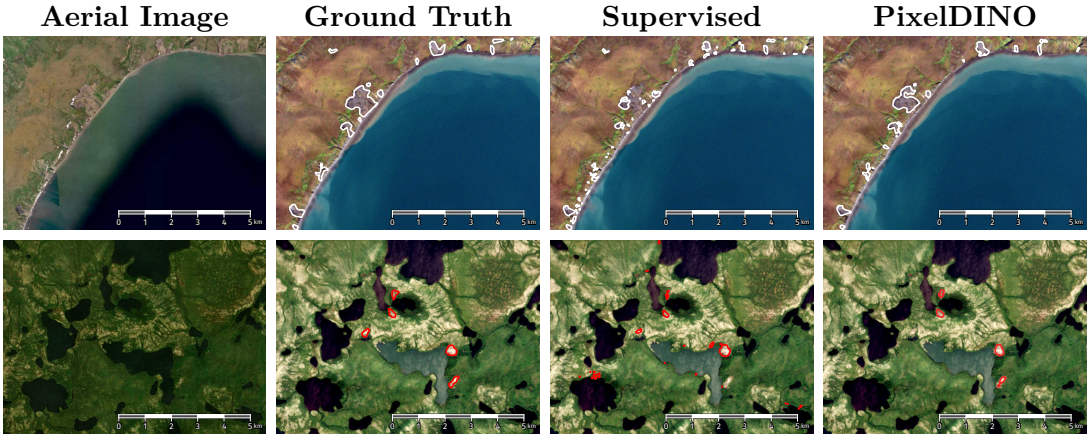


Figure 5.5: High resolution imagery (1st column), ground truth (2nd column), and prediction results for parts of the Herschel Island (top) and Lena (bottom) study sites for the Baseline+Aug (3rd column) and PixelDINO (4th column) training methods. Most prominent is the large reduction in false positives due to the semi-supervised training method. The visualisations in columns 2-4 are displayed on top of Sentinel-2 data from the test datasets, high resolution imagery in column 1 courtesy of Esri, Maxar, Earthstar Geographics, and the GIS User Community. Figure taken from [5].

For the Russian Arctic, RTS data was added for the Novaya Zemlya archipelago, the Gydan and Taimyr peninsulas, and the Lena river delta. These regions can be seen in figure 5.4. For testing spatial generalisation, the Herschel and Lena sites were set aside, as the Lena region is the only region far inland and Herschel is an island isolated from the other study regions.

As access to commercial PlanetScope data hinders reproducibility and complicates the acquisition of a large unlabelled dataset, this study was not conducted on the original PlanetScope data, but used Sentinel-2 multi-spectral data. While the resolution of Sentinel-2 is only 10 m as opposed to PlanetScope's roughly 3 m resolution, Sentinel-2 data is freely available and features 13 spectral bands.

The existing RTS footprints were re-projected to match the pixel grid of the Sentinel-2 imagery. Further, the auxiliary input data modalities of time-series trends and elevation information were removed to simplify the model pipeline. For the semi-supervised training methods, an unlabelled dataset was collected by arbitrarily selecting 83 Sentinel-2 grid tiles in regions of ice-rich permafrost, which were known or conjectured to contain RTSs. These Sentinel-2 scenes were processed in the same way as the scenes for the annotated regions, yielding compatible training datasets.

Having identified data efficiency as the most critical challenge in RTS detection, the simplest model from the previous study in section 5.1 was used, namely the UNet model [55]. As baselines, we first trained a UNet model on only the labelled data. In order to also quantify the effect of data augmentations, this model was trained both with and without data augmentations. Then, including the unlabelled dataset,

Herschel Island

Method	IoU	mIoU	F1	Precision	Recall
Baseline	19.8 ± 1.7	59.6 ± 0.9	33.0 ± 2.3	28.8 ± 3.0	39.4 ± 5.0
Baseline+Aug	22.9 ± 3.0	61.3 ± 1.5	37.2 ± 3.9	44.2 ± 7.5	32.3 ± 2.0
FixMatchSeg [156]	23.4 ± 0.8	61.5 ± 0.4	37.9 ± 1.1	34.1 ± 2.3	43.2 ± 4.5
Adversarial [163]	26.6 ± 3.9	63.2 ± 1.9	41.9 ± 4.9	60.0 ± 9.2	32.3 ± 3.1
PixelDINO	30.2 ± 2.7	65.0 ± 1.4	46.4 ± 3.2	52.7 ± 9.2	42.0 ± 3.0

Lena River

Method	IoU	mIoU	F1	Precision	Recall
Baseline	28.8 ± 4.0	64.3 ± 2.0	44.6 ± 5.0	52.8 ± 5.9	39.0 ± 6.0
Baseline+Aug	25.8 ± 10.2	62.8 ± 5.1	40.2 ± 13.0	69.4 ± 3.2	29.4 ± 12.5
FixMatchSeg [156]	32.4 ± 3.2	66.1 ± 1.6	48.8 ± 3.7	59.4 ± 2.7	41.6 ± 5.0
Adversarial [163]	25.1 ± 15.1	62.4 ± 7.5	38.2 ± 20.5	87.3 ± 7.5	26.8 ± 16.7
PixelDINO	39.5 ± 6.5	69.7 ± 3.3	56.4 ± 6.6	77.7 ± 6.3	44.5 ± 6.8

Table 5.1: Results of the Generalisation Study: Mean and Standard Deviation of 4 runs each for both the Herschel Island and Lena River test sites. (Values in %)

additional models were trained with semi-supervised training protocols. Besides the PixelDINO approach, two other semi-supervised training procedures for semantic segmentation were evaluated. FixMatchSeg [156] relies on semi-supervised learning from pseudo-labels. It inspired some ideas used in PixelDINO, such as the composition of weak and strong augmentations. The second approach, AdvSemiSeg [163], relies on adversarial learning by training a discriminator network to discern true class maps from predicted class maps. Once again, the models were evaluated using the IoU metric. The evaluation results and their standard deviations reported in A.5 are reproduced in table 5.1.

Surprisingly, the effect of data augmentation was not uniform across the two evaluated study regions. While data augmentation slightly improved performance for the Herschel Study site, it actually worsened the performance for the Lena site. This unclear trend suggests that the data augmentations might not always be as helpful for generalisation as they are generally assumed to be.

Comparing the supervised approaches to the semi-supervised approaches, it is clear to see that semi-supervised learning is beneficial for RTS detection. IoU scores generally improve when using semi-supervised learning with the exception of the adversarially trained model on the Lena region. What is more, the PixelDINO approach significantly outperforms the other semi-supervised learning approaches. This can be attributed to the richer semi-supervised training feedback given through the pseudo-classes used in the PixelDINO procedure. While the other methods have to rely on learning information in the highly sparse existing label space, PixelDINO can use the much richer pseudo-class space for its semi-supervised learning.

Figure 5.5 shows the prediction results of a supervised model and one trained with PixelDINO. Visually, the main improvement from PixelDINO appears to be the the reduction of false positive predictions. While the supervised model is overly sensitive in areas near bodies of water, this effect is greatly reduced when using PixelDINO. That suggests that models trained with PixelDINO appear less likely to predict the presence of RTS in regions without any RTS.

To sum up, PixelDINO is an effective method for semi-supervised semantic segmentation for tasks with sparse targets, such as RTS detection. The introduction of pseudo-classes allows for efficient consistency regularisation even in these settings, where methods relying on the original label space do not bring much improvement. Combining the concept of self-distillation with no labels [169] with ideas from existing semi-supervised semantic segmentation approaches [156] allows the model to learn pixel-wise features of high semantic value.

6 Conclusion & Outlook

This dissertation developed new deep learning methodology for automating some of the analytical tasks in polar remote sensing.

The impacts of global climate change on the polar regions were highlighted in chapter 1, motivating the importance of monitoring these particular regions. Chapter 2 then gave an overview over deep learning as the primary algorithmic tool of this dissertation and reviewed the ways deep learning is being used in remote sensing. The specific challenges of remote sensing in the polar regions were also discussed. Existing approaches to automatically map features in the polar regions were categorised and discussed in chapter 3.

Chapter 4 outlined the scientific contributions made in this dissertation regarding glacier calving front mapping. The HED-UNet model was designed to detect calving fronts in Antarctica by closely observing how a human approaches calving front detection, and designing a deep learning model according to these observations. Key design elements include the combination of segmentation with edge detection and the ability of the model to attend to different resolution levels. The second model, COBRA, takes this a step further and directly predicts the desired contour lines instead of pixel-wise masks. In the third study of this chapter, the possibilities of applying COBRA to a new region, namely Svalbard, were highlighted.

Chapter 5 discussed the scientific contributions regarding permafrost degradation mapping. The first study explored the overall feasibility of applying deep learning for this task, and identified spatial generalisation and limited labels as the main challenges. The second study proposed PixelDINO, a method using both labelled and unlabelled satellite images in a semi-supervised fashion. In this way, the performance of deep learning models for mapping retrogressive thaw slumps (RTSs) could be greatly improved.

The scientific insights gained over the course of this dissertation can be summarised as follows:

Importances of Edges For detecting glacier calving fronts, focusing on the edges instead of land and ocean areas is highly promising. Recent research appears to converge around this idea, supporting the assumption that this is indeed the way forward for calving front detection [1], [2], [8], [104].

Spatial Context Matters Glaciers can measure dozens of kilometres in size. This means that to fully understand such large-scale features, deep learning models need to be able to take into account a broad spatial context. Therefore, deep learning models designed for computer vision or other remote sensing tasks like mapping building footprints are often suboptimal in this application context. Instead, specific models need to be designed taking into account the need for spatial context information.

Task Representation Matters Many deep learning studies for remote sensing apply an existing model that was originally designed for a different task. In this way, implicit assumptions about the nature of the task are made. Taking a step back and finding the optimal way of computationally encoding a task can help rectifying such issues and improve model performance.

Efficient Representations for Uncertainty Quantification As the explicit contour representation of the COBRA model has shown, using more efficient data representations can help with quantifying model uncertainties. This is explained by the observation that a smaller data representation allows for less interaction terms between its components, facilitating uncertainty estimation.

Downstream Usefulness The presented deep learning models can automatically process large satellite imagery archives in order to compute multi-decadal time series of predictions for large study areas. The resulting data products are reliable enough for downstream use and contain valuable insights.

Quantity of Training Data Large amounts of training data are required to train robust deep learning models. While techniques such as semi-supervised learning can help alleviate this problem, a larger training dataset is often the easiest way to improve model performance.

Resolution Does Not Matter Much Intuitively, higher resolution imagery allows for more precise mapping of features such as calving fronts and RTS. In both applications, however, it turned out that working with lower resolution data like Landsat or Sentinel-2 imagery can yield comparable results to those obtained from higher resolution data like PlanetScope [2]. The reason might be that medium resolution sensors often have more spectral bands, and the fact that the spatial context that a model will take into account increases proportionally to the pixel spacing.

Consistency Without Pre-Defined Labels For semi-supervised semantic segmentation, it is possible to enforce consistency of the model predictions without using predefined labels. As explained in the PixelDINO study, the model can come up with its own segmentation classes. Forcing the model to make use of all these classes prevents output collapse and allows the model to learn meaningful features from the consistency learning feedback.

Figure 6.1: Conceptual timeline of deep learning studies for remote sensing in the polar regions. After establishing the feasibility of deep learning for a certain task, task-specific models are developed next. After a while, promising approaches are singled out and operationalised for providing large-scale predictions. Finally, the derived data can be used for downstream studies.

6.1 Conceptual Timeline of Polar Remote Sensing Research

Like all research, the contributions from this dissertation project do not stand for themselves, but constitute specific insights and building blocks towards a larger scientific goal. In this case, the penultimate goal is to foster a better understanding of the processes and dynamics happening in the glacial and periglacial environments in the Arctic and Antarctica. We will take a step back to look at the bigger picture in order to understand the current state of the field, as well as where it might be headed in the near future.

When looking at the numerous studies done in remote sensing for the polar regions, they can be divided roughly into the following four phases, where each phase builds upon the previous one, as symbolised in figure 6.1.

1. **Feasibility Studies** The first group of studies started by seeing the great potential of the deep learning models developed in computer vision. By applying the same concepts to polar remote sensing tasks, these studies evaluate the feasibility of deep learning for such use cases. For example, initial studies apply the UNet model [55] as a first baseline for many use cases. Feasibility studies generally confirmed the great potential of these models, as described in chapter 3. However, in this early stage, studies were usually confined to selected study areas and the models trained on small datasets. In order to scale up to more reliable model predictions on a pan-Arctic scale, larger datasets and more specialised models are needed as a next step.

2. **Specialised Studies** Encouraged by the positive results of the feasibility studies, this second group of studies started to develop deep learning models that were more tailored for the tasks of polar remote sensing. Seeing that different tasks require different approaches, models were adapted and improved for each task, like calving front detection, sea-ice charting, or permafrost mapping. One example for this is the move from segmentation to edge detectors, which was discussed in section 4.1. At the same time, a push towards larger datasets with greater spatial coverage is observed. These extensive datasets then serve as benchmarks for new methodological developments in the field. Four of the manuscripts contained within this dissertation belong to this second phase [1], [2], [4], [5].

3. Operationalisation At some point the model predictions become reliable enough for application-specific analyses. However, detailed knowledge of data processing toolchains and deep learning frameworks is needed to generate predictions from the trained models. Operationalisation studies fill the gap as an intermediary step between the methodological studies from phase 2 and the downstream studies in phase 4. These studies explore how to scale the deep learning models to large satellite image archives. By making the model predictions available in formats compatible with standard geographic information system (GIS) applications, they make it easy for polar scientists to use these data for downstream studies. The Svalbard study from this dissertation is part of this phase [3].

4. Downstream Studies While downstream studies are still in their early stages, it can be expected that exciting studies will soon leverage this wealth of automatically derived data. This will lead to a better understanding of the monitored processes on a pan-Arctic level. For this, the derived observational data products can be combined with other types of data like weather and climate data to identify and quantify the interdependencies between environmental factors and the actual reactions of the polar systems. Another promising application is the combination of the data products with physical modelling approaches to improve the accuracy of these physical models.

For calving front monitoring, the HED-UNet model has been operationalised for Antarctic glaciers. The resulting predictions are made easily available in an analysis-ready shapefile format in the *IceLines* data product [9]. The COBRA model has been deployed for the analysis of Svalbard's marine-terminating glaciers, with the resulting predictions being readily available for further analysis [3]. A comparable dataset is also available for calving fronts in Greenland [170]. Finally, first steps towards operationalisation are also taken for RTS detection [128].

6.2 Possibilities for Follow-Up Research

This section introduces some ideas how deep learning techniques could be used to further help polar research.

Temporal Reasoning The tasks approached in this dissertation project are motivated by the need to monitor dynamic processes in the polar regions. Therefore, they can be more easily detected and understood in a temporal context. However, current studies for deep learning in polar remote sensing only make predictions for single points in time. Methods like convolutional long short-term memory (ConvLSTM) models allow for a combination of vision methods with sequence models [171]. Applying such approaches for polar remote sensing could greatly benefit the tasks studied in this dissertation. Phenomena like the surging of glaciers [3], or the polycyclicity of RTS [172], could be better understood using models that have a concept of time.

Predictive Modelling Following the idea of using temporal context, models could not only use temporal information to better detect the monitored features, but even forecast their developments into the future. Recent deep learning methods from fields like video prediction [173] or weather forecasting [174] allow for the prediction of future imagery given a recent history. Adapting such approaches to forecast trends in glacier dynamics or the growth of RTS could greatly help polar research to estimate the developments in the polar regions for the future.

Physics-Aware Machine Learning Another approach for improving the performance of deep learning models for cold regions could be the inclusion of physical knowledge into the deep learning models. Sophisticated physical models like the Open Global Glacier Model (OGGM) [175] and CryoGrid [176] can numerically simulate the dynamics of glaciers and permafrost regions, respectively. The field of physics-aware machine learning [177] explores ways of integrating such physical models with deep learning. In this way, phyiscal process understanding can be combined with the flexible learning process of neural networks. Bolibar *et al.* [178] presented a first model combining neural networks and differential equations for modelling the dynamics of mountain glaciers. Such approaches may greatly improve models for deep learning in polar regions.

Exploiting Task Synergies So far, deep learning is only being used to tackle specific tasks in polar remote sensing. In these settings, complex deep learning models struggle with the limited amount of training data. As the various polar remote sensing tasks are analysing similar features and regions, a multi-task model for polar remote sensing might be able to exploit synergies between these tasks.

Causal Modelling A recent line of research tries to use deep learning to model and understand causal relationships between certain variables [179]. Data-driven analysis of the causal relationships between climate variables and cryosphere processes might uncover previously unknown connections and help quantify the effect of those already known.

6.3 Outlook

The polar regions are large areas experiencing heavy effects from global climate change. However, the full extent of the changes in these regions is hard to quantify. Remote sensing data has incredible potential for monitoring these regions and better understanding the underlying processes. With larger collections of satellite images becoming available to the public every year, these datasets are a treasure trove for polar research.

For the analysis of these data archives, innovative methods have been proposed in recent years, including the studies presented in this dissertation. By automating mapping and detection tasks for the polar regions, valuable insights can be extracted from these large datasets. By reducing terrabytes of information into key properties like the

positions of calving fronts or the footprints of retrogressive thaw slumps, it becomes much easier to study the fundamental processes.

In the near future, we can expect research to gain additional understanding of key processes in this data-driven way. For complex phenomena like glacier surges or the polycyclic growth of retrogressive thaw slumps, it will be tremendously helpful to not only look at single occurrences, but instead observe a large number of instances with the help of machine learning.

For deep learning to reach its full potential in polar research, it is of paramount importance that machine learning researchers and polar researchers keep working hand in hand. Only in this way can it be assured that the monitored targets are well captured by the models, the predictions are reliable and, finally, the derived data is helpful for polar research.

Bibliography

[1] K. Heidler, L. Mou, C. Baumhoer, A. Dietz, and X. X. Zhu, "HED-UNet: Combined segmentation and edge detection for monitoring the antarctic coastline," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022. DOI: 10.1109/TGRS.2021.3064606.

[2] K. Heidler, L. Mou, E. Loebel, M. Scheinert, S. Lefèvre, and X. X. Zhu, "A deep active contour model for delineating glacier calving fronts," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–12, 2023. DOI: 10.1109/TGRS.2023.3296539.

[3] T. Li, K. Heidler, L. Mou, Á. Ignéczi, X. X. Zhu, and J. L. Bamber, "A high-resolution calving front data product for marine-terminating glaciers in Svalbard," *Earth System Science Data*, vol. 16, no. 2, pp. 919–939, 2024. DOI: 10.5194/essd-16-919-2024.

[4] I. Nitze, K. Heidler, S. Barth, and G. Grosse, "Developing and testing a deep learning approach for mapping retrogressive thaw slumps," *Remote Sensing*, vol. 13, no. 21, p. 4294, 2021. DOI: 10.3390/rs13214294.

[5] K. Heidler, I. Nitze, G. Grosse, and X. X. Zhu, "PixelDINO: Semi-supervised semantic segmentation for detecting permafrost disturbances," *IEEE Transactions on Geoscience and Remote Sensing (in review)*, 2024. DOI: 10.48550/arXiv.2401.09271.

[6] M. Rantanen, A. Y. Karpechko, A. Lipponen, K. Nordling, O. Hyvärinen, K. Ruosteenoja, T. Vihma, and A. Laaksonen, "The Arctic has warmed nearly four times faster than the globe since 1979," *Communications Earth & Environment*, vol. 3, no. 1, pp. 1–10, 2022. DOI: 10.1038/s43247-022-00498-3.

[7] M. Casado, R. Hébert, D. Faranda, and A. Landais, "The quandary of detecting the signature of climate change in Antarctica," *Nature Climate Change*, vol. 13, no. 10, pp. 1082–1088, 2023. DOI: 10.1038/s41558-023-01791-5.

[8] D. Cheng, W. Hayes, E. Larour, Y. Mohajerani, M. Wood, I. Velicogna, and E. Rignot, "Calving Front Machine (CALFIN): Glacial termini dataset and automated deep learning extraction method for Greenland, 1972–2019," *The Cryosphere*, vol. 15, no. 3, pp. 1663–1675, 2021. DOI: 10.5194/tc-15-1663-2021.

[9] C. A. Baumhoer, A. J. Dietz, K. Heidler, and C. Kuenzer, "IceLines – A new data set of Antarctic ice shelf front positions," *Scientific Data*, vol. 10, no. 1, p. 138, 2023. DOI: 10.1038/s41597-023-02045-x.

[10] K. Calvin, D. Dasgupta, G. Krinner, *et al.*, "IPCC, 2023: Climate change 2023: Synthesis report. Contribution of working groups I, II and III to the sixth assessment report of the intergovernmental panel on climate change [Core writing team, H. Lee and J. Romero (eds.)]," Intergovernmental Panel on Climate Change (IPCC), Geneva, Switzerland, 2023. DOI: 10.59327/IPCC/AR6-9789291691647.

[11] D. J. Cavalieri and C. L. Parkinson, "Arctic sea ice variability and trends, 1979–2010," *The Cryosphere*, vol. 6, no. 4, pp. 881–889, 2012. DOI: 10.5194/tc-6-881-2012.

[12] S. Smith, V. Romanovsky, A. Lewkowicz, C. Burn, M. Allard, G. Clow, K. Yoshikawa, and J. Throop, "Thermal state of permafrost in North America: A contribution to the international polar year," *Permafrost and Periglacial Processes*, vol. 21, no. 2, pp. 117–135, 2010. DOI: 10.1002/ppp.690.

[13] D. M. Holland, K. W. Nicholls, and A. Basinski, "The Southern Ocean and its interaction with the Antarctic Ice Sheet," *Science*, vol. 367, no. 6484, pp. 1326–1330, 2020. DOI: 10.1126/science.aaz5491.

[14] N. R. Golledge, E. D. Keller, N. Gomez, K. A. Naughten, J. Bernales, L. D. Trusel, and T. L. Edwards, "Global environmental consequences of twenty-first-century ice-sheet melt," *Nature*, vol. 566, no. 7742, pp. 65–72, 2019. DOI: 10.1038/s41586-019-0889-9.

[15] A. Shepherd, E. Ivins, E. Rignot, *et al.*, "Mass balance of the Antarctic Ice Sheet from 1992 to 2017," *Nature*, vol. 558, no. 7709, pp. 219–222, 2018. DOI: 10.1038/s41586-018-0179-y.

[16] A. Shepherd, E. Ivins, E. Rignot, *et al.*, "Mass balance of the Greenland Ice Sheet from 1992 to 2018," *Nature*, vol. 579, no. 7798, pp. 233–239, 2020. DOI: 10.1038/s41586-019-1855-2.

[17] S. Doney, M. Ruckelshaus, J. E. Duffy, *et al.*, "Climate change impacts on marine ecosystems.," *Annual review of marine science*, 2012. DOI: 10.1146/ANNUREV-MARINE-041911-111611.

[18] P. Zhang, Y. Wu, G. Chen, and Y. Yu, "North American cold events following sudden stratospheric warming in the presence of low Barents-Kara Sea sea ice," *Environmental Research Letters*, vol. 15, no. 12, p. 124 017, 2020. DOI: 10.1088/1748-9326/abc215.

[19] I. V. Polyakov, A. V. Pnyushkov, M. B. Alkire, *et al.*, "Greater role for Atlantic inflows on sea-ice loss in the Eurasian Basin of the Arctic Ocean," *Science*, vol. 356, no. 6335, pp. 285–291, 2017. DOI: 10.1126/science.aai8204.

[20] M.-L. Timmermans and J. Marshall, "Understanding Arctic Ocean Circulation: A Review of Ocean Dynamics in a Changing Climate," *Journal of Geophysical Research: Oceans*, vol. 125, no. 4, 2020. DOI: 10.1029/2018JC014378.

[21] U. of Tasmania. "Antarctic researchers' rare view of an Ice Shelf calving - Institute for Marine and Antarctic Studies," Institute for Marine and Antarctic Studies - University of Tasmania, Australia. (2019), [Online]. Available: https://www.imas.utas.edu.au/news/news-items/antarctic-researchers-rare-view-of-an-ice-shelf-calving (visited on 01/15/2024).

[22] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, "Google Earth Engine: Planetary-scale geospatial analysis for everyone," *Remote Sensing of Environment*, 2017. DOI: 10.1016/j.rse.2017.06.031.

[23] Intergovernmental Panel On Climate Change, *Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, 1st ed. Cambridge University Press, 2023, ISBN: 978-1-00-915789-6. DOI: 10.1017/9781009157896.

[24] I. Nitze, G. Grosse, B. M. Jones, V. E. Romanovsky, and J. Boike, "Remote sensing quantifies widespread abundance of permafrost region disturbances across the Arctic and Subarctic," *Nature Communications*, vol. 9, no. 1, pp. 1–11, 2018. DOI: 10.1038/s41467-018-07663-3.

[25] Y. Shur, M. T. Jorgenson, and M. Z. Kanevskiy, "Permafrost," in *Encyclopedia of Snow, Ice and Glaciers*, ser. Encyclopedia of Earth Sciences Series, Dordrecht: Springer Netherlands, 2011, pp. 841–848, ISBN: 978-90-481-2642-2. DOI: 10.1007/978-90-481-2642-2_400.

[26] J. Obu, "How much of the Earth's surface is underlain by permafrost?" *Journal of Geophysical Research: Earth Surface*, vol. 126, no. 5, e2021JF006123, 2021. DOI: 10.1029/2021JF006123.

[27] M. Liew, X. Ji, M. Xiao, L. Farquharson, D. Nicolsky, V. Romanovsky, M. Bray, X. Zhang, and C. McComb, "Synthesis of physical processes of permafrost degradation and geophysical and geomechanical properties of permafrost," *Cold Regions Science and Technology*, vol. 198, p. 103 522, 2022. DOI: 10.1016/j.coldregions.2022.103522.

[28] V. P. Melnikov, V. I. Osipov, A. V. Brouchkov, *et al.*, "Climate warming and permafrost thaw in the Russian Arctic: Potential economic impacts on public infrastructure by 2050," *Natural Hazards*, vol. 112, no. 1, pp. 231–251, 2022. DOI: 10.1007/s11069-021-05179-6.

[29] E. a. G. Schuur, A. D. McGuire, C. Schädel, *et al.*, "Climate change and the permafrost carbon feedback," *Nature*, vol. 520, no. 7546, pp. 171–179, 2015. DOI: 10.1038/nature14338.

[30] A. Bartsch, S. Westemann, T. Strozzi, and F. M. Seifert, "CCI+ PHASE 1 – New ECVS Permafrost: Product validation and algorithm selection report," European Space Agency (ESA), D2.1, 2021.

[31] A. D. Parsekian, R. H. Chen, R. J. Michaelides, *et al.*, "Validation of permafrost active layer estimates from airborne SAR observations," *Remote Sensing*, vol. 13, no. 15, p. 2876, 2021. DOI: 10.3390/rs13152876.

[32] A. Vieli, M. Funk, and H. Blatter, "Flow dynamics of tidewater glaciers: A numerical modelling approach," *Journal of Glaciology*, vol. 47, no. 159, pp. 595–606, 2001. DOI: 10.3189/172756501781831747.

[33] A. Levermann, T. Albrecht, R. Winkelmann, M. A. Martin, M. Haseloff, and I. Joughin, "Kinematic first-order calving law implies potential for abrupt ice-shelf retreat," *The Cryosphere*, vol. 6, no. 2, pp. 273–286, 2012. DOI: 10.5194/tc-6-273-2012.

[34] M. Morlighem, J. Bondzio, H. Seroussi, E. Rignot, E. Larour, A. Humbert, and S. Rebuffi, "Modeling of Store Gletscher's calving dynamics, West Greenland, in response to ocean thermal forcing," *Geophysical Research Letters*, vol. 43, no. 6, pp. 2659–2666, 2016. DOI: 10.1002/2016GL067695.

[35] Y. Choi, M. Morlighem, M. Wood, and J. H. Bondzio, "Comparison of four calving laws to model Greenland outlet glaciers," *The Cryosphere*, vol. 12, no. 12, pp. 3735–3746, 2018. DOI: 10.5194/tc-12-3735-2018.

[36] J. A. Wilner, M. Morlighem, and G. Cheng, "Evaluation of four calving laws for Antarctic ice shelves," *The Cryosphere*, vol. 17, no. 11, pp. 4889–4901, 2023. DOI: 10.5194/tc-17-4889-2023.

[37] E. Buch, M. S. Madsen, J. She, M. Stendel, O. K. Leth, A. M. Fjæraa, and M. Rattenborg, "Arctic in situ data avilability," European Environment Agency, Kobenhavn, Denmark, 2.1, 2019.

[38] C. Gabarró, N. Hughes, J. Wilkinson, *et al.*, "Improving satellite-based monitoring of the polar regions: Identification of research and capacity gaps," in *Frontiers in Remote Sensing*, vol. 4, 2023, p. 952 091. DOI: 10.3389/frsen.2023.952091.

[39] A. Bartsch, T. Strozzi, and I. Nitze, "Permafrost Monitoring from Space," *Surveys in Geophysics*, 2023. DOI: 10.1007/s10712-023-09770-3.

[40] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016, ISBN: 978-0-262-03561-3.

[41] C. M. Bishop, *Pattern recognition and machine learning* (Information science and statistics). New York: Springer, 2006, 738 pp., ISBN: 978-0-387-31073-2.

[42] K. P. Murphy, *Machine learning: a probabilistic perspective* (Adaptive computation and machine learning series). Cambridge, MA: MIT Press, 2012, 1067 pp., ISBN: 978-0-262-01802-9.

[43] K. Hornik, "Approximation capabilities of muitilayer feedforward networks," *Neural Networks*, vol. 4, p. 7, 1991.

[44] B. Hanin, "Universal function approximation by deep neural nets with bounded width and ReLU activations," 2017.

[45] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *3rd international conference on learning representations, ICLR*, 2015.

[46] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," 2022.

[47] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 8–36, 2017. DOI: 10.1109/MGRS.2017.2762307.

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. 2016 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 770–778, 2016. DOI: 10.1109/CVPR.2016.90.

[49] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[50] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018. DOI: 10.1109/TPAMI.2017.2699184.

[51] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, vol. 25, Curran Associates, Inc., 2012, pp. 1097–1105.

[52] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," presented at the International Conference on Learning Representations, 2020.

[53] M. Castelluccio, G. Poggi, C. Sansone, and L. Verdoliva, "Land Use Classification in Remote Sensing Images by Convolutional Neural Networks," *ArXiv*, 2015.

[54] G. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3965–3981, 2017. DOI: 10.1109/TGRS.2017.2685945.

[55] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pp. 234–241. DOI: 10.1007/978-3-319-24574-4_28.

[56] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes dataset for semantic urban scene understanding," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA: IEEE, 2016, pp. 3213–3223, ISBN: 978-1-4673-8851-1. DOI: 10.1109/CVPR.2016.350.

[57] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, "Deep learning classification of land cover and crop types using remote sensing data," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 5, pp. 778–782, 2017. DOI: 10.1109/LGRS.2017.2681128.

[58] J. Yang, J. Guo, H. Yue, Z. Liu, H. Hu, and K. Li, "CDnet: CNN-based cloud detection for remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 8, pp. 6195–6211, 2019. DOI: 10.1109/TGRS.2019.2904868.

[59] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *CoRR*, vol. abs/1804.02767, 2018.

[60] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Computer Vision – ECCV 2014*, ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2014, pp. 740–755, ISBN: 978-3-319-10602-1. DOI: 10.1007/978-3-319-10602-1_48.

[61] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 117, pp. 11–28, 2016. DOI: 10.1016/j.isprsjprs.2016.03.014.

[62] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988. DOI: 10.1109/ICCV.2017.322.

[63] G. Liu, B. Peng, T. Liu, *et al.*, "Fine-grained building roof instance segmentation based on domain adapted pretraining and composite dual-backbone," *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, pp. 670–673, 2023. DOI: 10.1109/IGARSS52108.2023.10281999.

[64] L. Mou and X. X. Zhu, "Vehicle instance segmentation from aerial image and video using a multitask learning residual fully convolutional network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 11, pp. 6699–6711, 2018. DOI: 10.1109/TGRS.2018.2841808.

[65] J. Campbell and R. Wynne, *Introduction to remote sensing*, Fifth edition. Guilford Publications, 2011, ISBN: 978-1-60918-177-2.

[66] J. H. Meeus, *Astronomical algorithms*. Willmann-Bell, Incorporated, 1991, ISBN: 978-0-943396-35-4.

[67] "Landsat 9," Reston, VA, Report 2019-3008, 2019, p. 2. DOI: 10.3133/fs20193008.

[68] F. Languille, C. Déchoz, A. Gaudel, D. Greslou, F. de Lussy, T. Trémas, and V. Poulain, "Sentinel-2 geometric image quality commissioning: First results," in *Image and Signal Processing for Remote Sensing XXI*, vol. 9643, SPIE, 2015, pp. 61–73. DOI: 10.1117/12.2194339.

[69] I. Yalcin, S. Kocaman, S. Saunier, and C. Albinet, "Radiometric quality assessment for Maxar HD imagery," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLIII-B3-2021, pp. 797–804, 2021. DOI: 10.5194/isprs-archives-XLIII-B3-2021-797-2021.

[70] D. P. Roy, H. Huang, R. Houborg, and V. S. Martins, "A global analysis of the temporal availability of PlanetScope high spatial resolution multi-spectral imagery," *Remote Sensing of Environment*, vol. 264, p. 112 586, 2021. DOI: 10.1016/j.rse.2021.112586.

[71] C. Burn, *The polar night*. Aurora Research Institute, 1996.

[72] M. O. Source, M. McFarland, D. Emanuele, D. Morris, and T. Augspurger, *Microsoft/PlanetaryComputer: October 2022*, version 2022.10.28, Zenodo, 2022. DOI: 10.5281/zenodo.7261897.

[73] M. Bourbigot, H. Johnsen, and R. Piantanida, *Sentinel-1 Product Definition*, version 2.7, 2016, S1-RS-MDA-52-7440.

[74] J. Mouginot, E. Rignot, B. Scheuchl, and R. Millan, "Comprehensive annual ice sheet velocity mapping using Landsat-8, Sentinel-1, and RADARSAT-2 data," *Remote Sensing*, vol. 9, no. 4, p. 364, 2017. DOI: 10.3390/rs9040364.

[75] A. Bartsch, G. Pointner, H. Bergstedt, B. Widhalm, A. Wendleder, and A. Roth, "Utility of polarizations available from Sentinel-1 for tundra mapping," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, 2021, pp. 1452–1455. DOI: 10.1109/IGARSS47720.2021.9553993.

[76] C. A. Baumhoer, A. J. Dietz, C. Kneisel, and C. Kuenzer, "Automated extraction of antarctic glacier and ice shelf fronts from sentinel-1 imagery using deep learning," *Remote Sensing*, vol. 11, no. 21, p. 2529, 2019. DOI: 10.3390/rs11212529.

[77] D. I. Benn, C. R. Warren, and R. H. Mottram, "Calving processes and the dynamics of calving glaciers," *Earth-Science Reviews*, vol. 82, no. 3, pp. 143–179, 2007. DOI: 10.1016/j.earscirev.2007.02.002.

[78] B. Kamb, C. F. Raymond, W. D. Harrison, H. Engelhardt, K. A. Echelmeyer, N. Humphrey, M. M. Brugman, and T. Pfeffer, "Glacier surge mechanism: 1982-1983 surge of Variegated Glacier, Alaska," *Science*, vol. 227, no. 4686, pp. 469–479, 1985. DOI: 10.1126/science.227.4686.469.

[79] H. Liu and K. C. Jezek, "A complete high-resolution coastline of Antarctica extracted from orthorectified Radarsat SAR Imagery," *Photogrammetric Engineering & Remote Sensing*, vol. 70, no. 5, pp. 605–616, 2004. DOI: 10.14358/PERS.70.5.605.

[80] Y. Liu, J. C. Moore, X. Cheng, R. M. Gladstone, J. N. Bassis, H. Liu, J. Wen, and F. Hui, "Ocean-driven thinning enhances iceberg calving and retreat of Antarctic ice shelves," *Proceedings of the National Academy of Sciences*, vol. 112, no. 11, pp. 3263–3268, 2015. DOI: 10.1073/pnas.1415137112.

[81] M. Schmitt, G. Baier, and X. X. Zhu, "Potential of nonlocally filtered pursuit monostatic TanDEM-X data for coastline detection," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 148, pp. 130–141, 2019. DOI: 10.1016/j.isprsjprs.2018.12.007.

[82] J.-S. Lee and I. Jurkevich, "Coastline detection and tracing In SAR images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 28, no. 4, pp. 662–668, 1990. DOI: 10.1109/TGRS.1990.572976.

[83] L. Krieger and D. Floricioiu, "Automatic calving front delineation on TerraSAR-X and Sentinel-1 SAR imagery," in *Proc. 2017 IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, 2017, pp. 2817–2820, ISBN: 978-1-5090-4951-6. DOI: 10.1109/IGARSS.2017.8127584.

[84] D. Wang and X. Liu, "Coastline extraction from SAR images using robust ridge tracing," *Marine Geodesy*, vol. 42, no. 3, pp. 286–315, 2019. DOI: 10.1080/01490419.2019.1583147.

[85] M. Modava and G. Akbarizadeh, "Coastline extraction from SAR images using spatial fuzzy clustering and the active contour method," *International Journal of Remote Sensing*, vol. 38, no. 2, pp. 355–370, 2017. DOI: 10.1080/01431161.2016.1266104.

[86] M. Modava, G. Akbarizadeh, and M. Soroosh, "Integration of spectral histogram and level set for coastline detection in SAR images," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 55, no. 2, pp. 810–819, 2019. DOI: 10.1109/TAES.2018.2865120.

[87] C. Liu, Y. Xiao, and J. Yang, "A coastline detection method in polarimetric SAR images mixing the region-based and edge-based active contour models," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3735–3747, 2017. DOI: 10.1109/TGRS.2017.2679112.

[88] T. Klinger, M. Ziems, C. Heipke, H. W. Schenke, and N. Ott, "Antarctic coastline detection using Snakes," *Photogrammetrie - Fernerkundung - Geoinformation*, vol. 2011, no. 6, pp. 421–434, 2011. DOI: 10.1127/1432-8364/2011/0095.

[89] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017. DOI: 10.1109/TPAMI.2016.2644615.

[90] X. Yuan, J. Shi, and L. Gu, "A review of deep learning methods for semantic segmentation of remote sensing imagery," *Expert Systems with Applications*, vol. 169, p. 114 417, 2021. DOI: 10.1016/j.eswa.2020.114417.

[91] P. Shamsolmoali, M. Zareapoor, R. Wang, H. Zhou, and J. Yang, "A novel deep structure U-Net for sea-land segmentation in remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 9, pp. 3219–3232, 2019. DOI: 10.1109/JSTARS.2019.2925841.

[92] R. Li, W. Liu, L. Yang, S. Sun, W. Hu, F. Zhang, and W. Li, "DeepUNet: A deep fully convolutional network for pixel-level sea-land segmentation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 11, pp. 3954–3962, 2018. DOI: 10.1109/JSTARS.2018.2833382.

[93] Z. Chu, T. Tian, R. Feng, and L. Wang, "Sea-land segmentation with Res-UNet and fully connected CRF," in *Proc. 2019 IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, 2019, pp. 3840–3843, ISBN: 978-1-5386-9154-0. DOI: 10.1109/IGARSS.2019.8900625.

[94] Y. Mohajerani, M. Wood, I. Velicogna, and E. Rignot, "Detection of glacier calving margins with convolutional neural networks: A case study," *Remote Sensing*, vol. 11, no. 1, p. 74, 2019. DOI: 10.3390/rs11010074.

[95] E. Zhang, L. Liu, and L. Huang, "Automatically delineating the calving front of Jakobshavn Isbræ from multitemporal TerraSAR-X images: A deep learning approach," *The Cryosphere*, vol. 13, no. 6, pp. 1729–1741, 2019. DOI: 10.5194/tc-13-1729-2019.

[96] M. Periyasamy, A. Davari, T. Seehaus, M. Braun, A. Maier, and V. Christlein, "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*, vol. 15, pp. 1712–1723, 2022. DOI: 10.1109/JSTARS.2022.3148033.

[97] E. Loebel, M. Scheinert, M. Horwath, K. Heidler, J. Christmann, L. D. Phan, A. Humbert, and X. X. Zhu, "Extracting glacier calving fronts by deep learning: The benefit of multispectral, topographic, and textural input features," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022. DOI: 10.1109/TGRS.2022.3208454.

[98] M. Holzmann, A. Davari, T. Seehaus, M. Braun, A. Maier, and V. Christlein, "Glacier calving front segmentation using attention U-Net," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, 2021, pp. 3483–3486. DOI: 10.1109/IGARSS47720.2021.9555067.

[99] A. Davari, C. Baller, T. Seehaus, M. Braun, A. Maier, and V. Christlein, "Pixelwise distance regression for glacier calving front detection and segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–10, 2022. DOI: 10.1109/TGRS.2022.3158591.

[100] F. Wu, N. Gourmelon, T. Seehaus, J. Zhang, M. Braun, A. Maier, and V. Christlein, "AMD-HookNet for glacier front segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–12, 2023. DOI: 10.1109/TGRS.2023.3245419.

[101] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI: IEEE, 2017, pp. 1800–1807. DOI: 10.1109/CVPR.2017.195.

[102] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Computer Vision – ECCV 2018*, vol. 11211, Cham: Springer International Publishing, 2018, pp. 833–851. DOI: 10.1007/978-3-030-01234-2_49.

[103] E. Zhang, L. Liu, L. Huang, and K. S. Ng, "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*, vol. 254, p. 112 265, 2021. DOI: 10.1016/j.rse.2020.112265.

[104] N. Gourmelon, T. Seehaus, M. Braun, A. Maier, and V. Christlein, "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*, vol. 14, no. 9, pp. 4287–4313, 2022. DOI: 10.5194/essd-14-4287-2022.

[105] C. R. Burn, "The thermal regime of a retrogressive thaw slump near Mayo, Yukon Territory," *Canadian Journal of Earth Sciences*, vol. 37, no. 7, pp. 967–981, 2000. DOI: 10.1139/e00-017.

[106] H. Lantuit and W. H. Pollard, "Temporal stereophotogrammetric analysis of retrogressive thaw slumps on Herschel Island, Yukon Territory," *Natural Hazards and Earth System Sciences*, vol. 5, no. 3, pp. 413–423, 2005. DOI: 10.5194/nhess-5-413-2005.

[107] M. R. Turetsky, B. W. Abbott, M. C. Jones, *et al.*, "Carbon release through abrupt permafrost thaw," *Nature Geoscience*, vol. 13, no. 2, pp. 138–143, 2020. DOI: 10.1038/s41561-019-0526-0.

[108] C. Burn and A. Lewkowicz, "Canadian landform examples - 17 retrogressive thaw slumps," *Canadian Geographies / Géographies canadiennes*, vol. 34, no. 3, pp. 273–276, 1990. DOI: 10.1111/j.1541-0064.1990.tb01092.x.

[109] S. V. Kokelj, T. C. Lantz, J. Tunnicliffe, R. Segal, and D. Lacelle, "Climate-driven thaw of permafrost preserved glacial landscapes, northwestern Canada," *Geology*, vol. 45, no. 4, pp. 371–374, 2017. DOI: 10.1130/G38626.1.

[110] T. C. Lantz and S. V. Kokelj, "Increasing rates of retrogressive thaw slump activity in the Mackenzie Delta region, N.W.T., Canada," *Geophysical Research Letters*, vol. 35, no. 6, 2008. DOI: 10.1029/2007GL032433.

[111] H. Lantuit and W. H. Pollard, "Fifty years of coastal erosion and retrogressive thaw slump activity on Herschel Island, southern Beaufort Sea, Yukon Territory, Canada," *Geomorphology*, Paraglacial Geomorphology: Processes and Paraglacial Context, vol. 95, no. 1, pp. 84–102, 2008. DOI: 10.1016/j.geomorph.2006.07.040.

[112] J. Obu, H. Lantuit, G. Grosse, F. Günther, T. Sachs, V. Helm, and M. Fritz, "Coastal erosion and mass wasting along the Canadian Beaufort Sea based on annual airborne LiDAR elevation data," *Geomorphology*, Permafrost and periglacial research from coasts to mountains, vol. 293, pp. 331–346, 2017. DOI: 10.1016/j.geomorph.2016.02.014.

[113] D. K. Swanson and M. Nolan, "Growth of retrogressive thaw slumps in the Noatak Valley, Alaska, 2010–2016, measured by airborne photogrammetry," *Remote Sensing*, vol. 10, no. 7, p. 983, 2018. DOI: 10.3390/rs10070983.

[114] J. Van der Sluijs, S. V. Kokelj, R. H. Fraser, J. Tunnicliffe, and D. Lacelle, "Permafrost terrain dynamics and infrastructure impacts revealed by UAV photogrammetry and thermal imaging," *Remote Sensing*, vol. 10, no. 11, p. 1734, 2018. DOI: 10.3390/rs10111734.

[115] E. P. Crist and R. C. Cicone, "A physically-based transformation of Thematic Mapper data—The TM Tasseled Cap," *IEEE Transactions on Geoscience and Remote Sensing*, vol. GE-22, no. 3, pp. 256–263, 1984. DOI: 10.1109/TGRS.1984.350619.

[116] A. Brooker, R. H. Fraser, I. Olthof, S. V. Kokelj, and D. Lacelle, "Mapping the activity and evolution of retrogressive thaw slumps by tasselled cap trend analysis of a landsat satellite image stack," *Permafrost and Periglacial Processes*, vol. 25, no. 4, pp. 243–256, 2014. DOI: 10.1002/ppp.1819.

[117] R. H. Fraser, I. Olthof, S. V. Kokelj, T. C. Lantz, D. Lacelle, A. Brooker, S. Wolfe, and S. Schwarz, "Detecting landscape changes in high latitude environments using Landsat trend analysis: 1. Visualization," *Remote Sensing*, vol. 6, no. 11, pp. 11 533–11 557, 2014. DOI: 10.3390/rs61111533.

[118] D. Lacelle, A. Brooker, R. H. Fraser, and S. V. Kokelj, "Distribution and growth of thaw slumps in the Richardson Mountains–Peel Plateau region, northwestern Canada," *Geomorphology*, vol. 235, pp. 40–51, 2015. DOI: 10.1016/j.geomorph.2015.01.024.

[119] P. Bernhard, S. Zwieback, S. Leinss, and I. Hajnsek, "Mapping retrogressive thaw slumps using single-pass TanDEM-X observations," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 3263–3280, 2020. DOI: 10.1109/JSTARS.2020.3000648.

[120] L. Huang, L. Liu, L. Jiang, and T. Zhang, "Automatic mapping of Thermokarst landforms from remote sensing Images using deep learning: A case study in the northeastern Tibetan Plateau," *Remote Sensing*, vol. 10, no. 12, p. 2067, 2018. DOI: 10.3390/rs10122067.

[121] L. Huang, J. Luo, Z. Lin, F. Niu, and L. Liu, "Using deep learning to map retrogressive thaw slumps in the Beiluhe region (Tibetan Plateau) from CubeSat images," *Remote Sensing of Environment*, vol. 237, p. 111 534, 2020. DOI: 10.1016/j.rse.2019.111534.

[122] C. Witharana, M. R. Udawalpola, A. K. Liljedahl, M. K. W. Jones, B. M. Jones, A. Hasan, D. Joshi, and E. Manos, "Automated detection of retrogressive thaw slumps in the High Arctic using high-resolution satellite imagery," *Remote Sensing*, vol. 14, no. 17, p. 4132, 2022. DOI: 10.3390/rs14174132.

[123] Y. Yang, B. M. Rogers, G. Fiske, J. Watts, S. Potter, T. Windholz, A. Mullen, I. Nitze, and S. M. Natali, "Mapping retrogressive thaw slumps using deep neural networks," *Remote Sensing of Environment*, vol. 288, p. 113 495, 2023. DOI: 10.1016/j.rse.2023.113495.

[124] A. Runge, I. Nitze, and G. Grosse, "Remote sensing annual dynamics of rapid permafrost thaw disturbances with LandTrendr," *Remote Sensing of Environment*, vol. 268, p. 112 752, 2022. DOI: 10.1016/j.rse.2021.112752.

[125] L. Huang, L. Liu, J. Luo, Z. Lin, and F. Niu, "Automatically quantifying evolution of retrogressive thaw slumps in Beiluhe (Tibetan Plateau) from multi-temporal CubeSat images," *International Journal of Applied Earth Observation and Geoinformation*, vol. 102, p. 102 399, 2021.

[126] Z. Xia, L. Huang, C. Fan, S. Jia, Z. Lin, L. Liu, J. Luo, F. Niu, and T. Zhang, "Retrogressive thaw slumps along the Qinghai–Tibet Engineering Corridor: A comprehensive inventory and their distribution characteristics," *Earth System Science Data*, vol. 14, no. 9, pp. 3875–3887, 2022. DOI: 10.5194/essd-14-3875-2022.

[127] L. Huang, T. C. Lantz, R. H. Fraser, K. F. Tiampo, M. J. Willis, and K. Schaefer, "Accuracy, efficiency, and transferability of a deep learning model for mapping retrogressive thaw slumps across the Canadian Arctic," *Remote Sensing*, vol. 14, no. 12, p. 2747, 2022. DOI: 10.3390/rs14122747.

[128] L. Huang, M. J. Willis, G. Li, T. C. Lantz, K. Schaefer, E. Wig, G. Cao, and K. F. Tiampo, "Identifying active retrogressive thaw slumps from ArcticDEM," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 205, pp. 301–316, 2023. DOI: 10.1016/j.isprsjprs.2023.10.008.

[129] C. Porter, P. Morin, I. Howat, *et al.*, *ArcticDEM, Version 3*, Harvard Dataverse, 2022. DOI: 10.7910/DVN/OHHUKH.

[130] A. Bochkovskiy, C.-Y. Wang, and H. Liao, "YOLOv4: Optimal speed and accuracy of object detection," *ArXiv*, 2020.

[131] G. Grosse, B. Jones, and C. Arp, "Thermokarst lakes, drainage, and drained basins," in *Treatise on Geomorphology*, San Diego: Academic Press, 2013, pp. 325–353, ISBN: 978-0-08-088522-3. DOI: 10.1016/B978-0-12-374739-6.00216-5.

[132] L. Hughes-Allen, F. Bouchard, A. Séjourné, G. Fougeron, and E. Léger, "Automated identification of Thermokarst lakes using machine learning in the ice-rich permafrost landscape of central Yakutia (Eastern Siberia)," *Remote Sensing*, vol. 15, no. 5, p. 1226, 2023. DOI: 10.3390/rs15051226.

[133] R. F. Black, "Periglacial features indicative of permafrost: Ice and soil wedges," *Quaternary Research*, vol. 6, no. 1, pp. 3–26, 1976. DOI: 10.1016/0033-5894(76)90037-5.

[134] T. Rettelbach, M. Langer, I. Nitze, B. Jones, V. Helm, J.-C. Freytag, and G. Grosse, "A quantitative graph-based approach to monitoring ice-wedge trough dynamics in polygonal permafrost landscapes," *Remote Sensing*, vol. 13, no. 16, p. 3098, 2021. DOI: 10.3390/rs13163098.

[135] C. J. Abolt, M. H. Young, A. L. Atchley, and C. J. Wilson, "Brief communication: Rapid machine-learning-based extraction and measurement of ice wedge polygons in high-resolution digital elevation models," *The Cryosphere*, vol. 13, no. 1, pp. 237–245, 2019. DOI: 10.5194/tc-13-237-2019.

[136] C. M. Gibson, L. E. Chasmer, D. K. Thompson, W. L. Quinton, M. D. Flannigan, and D. Olefeldt, "Wildfire as a major driver of recent permafrost thaw in boreal peatlands," *Nature Communications*, vol. 9, no. 1, p. 3041, 2018. DOI: 10.1038/s41467-018-05457-1.

[137] K. Heidler, L. Mou, C. Baumhoer, A. Dietz, and X. X. Zhu, "HED-UNet: A multi-scale framework for simultaneous segmentation and edge detection," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, 2021, pp. 3037–3040. DOI: 10.1109/IGARSS47720.2021.9553585.

[138] S. Xie and Z. Tu, "Holistically-nested edge detection," in *IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 1395–1403. DOI: 10.1109/ICCV.2015.164.

[139] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30*, Long Beach, CA, USA, 2017, pp. 5998–6008.

[140] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 4905–4913, ISBN: 978-1-5108-3881-9.

[141] J. Avbelj, R. Muller, and R. Bamler, "A metric for polygon comparison and building extraction evaluation," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 1, pp. 170–174, 2015. DOI: 10.1109/LGRS.2014.2330695.

[142] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International Journal of Computer Vision*, vol. 1, no. 4, pp. 321–331, 1988. DOI: 10.1007/BF00133570.

[143] S. Peng, W. Jiang, H. Pi, X. Li, H. Bao, and X. Zhou, "Deep snake for real-time instance segmentation," in *2020 IEEE/CVF conference on computer vision and pattern recognition, CVPR 2020, seattle, WA, USA, june 13-19, 2020*, IEEE, 2020, pp. 8530–8539. DOI: 10.1109/CVPR42600.2020.00856.

[144] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, *The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results*.

[145] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978. DOI: 10.1109/TASSP.1978.1163055.

[146] M. Cuturi and M. Blondel, "Soft-DTW: A Differentiable Loss Function for Time-Series," in *Proceedings of the 34th International Conference on Machine Learning*, PMLR, 2017, pp. 894–903.

[147] Z. Liu, J. H. Liew, X. Chen, and J. Feng, "DANCE: A deep attentive contour model for efficient instance segmentation," in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA: IEEE, 2021, pp. 345–354. DOI: 10.1109/WACV48630.2021.00039.

[148] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ser. ICML'17, Sydney, NSW, Australia: JMLR.org, 2017, pp. 1321–1330.

[149] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *Proceedings of The 33rd International Conference on Machine Learning*, PMLR, 2016, pp. 1050–1059.

[150] G. Moholdt, J. Maton, M. Majerska, and J. Kohler, *Annual frontlines of marine-terminating glaciers on Svalbard*, npolar.no, 2021. DOI: 10.21334/NPOLAR.2021.D60A919A.

[151] I. Nitze, J. Van der Sluijs, S. Barth, *et al.*, "An experiment to compare digitized labels of retrogressive thaw slumps by domain experts," presented at the 6th European Conference on Permafrost, Puigcerdà, Spain, 2023.

[152] I. Nitze, K. Heidler, S. Barth, and G. Grosse, "Deep learning for mapping retrogressive thaw slumps across the Arctic," in *16th International Circumpolar Remote Sensing Symposium*, Fairbanks, AK, USA, 2022.

[153] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 6, pp. 1856–1867, 2020. DOI: 10.1109/TMI.2019.2959609.

[154] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, 2020. DOI: 10.1109/TPAMI.2018.2858826.

[155] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, vol. 3, Atlanta, 2013, p. 896.

[156] P. Upretee and B. Khanal, "FixMatchSeg: Fixing FixMatch for semi-supervised semantic segmentation," 2022. DOI: 10.48550/arXiv.2208.00400.

[157] V. Verma, K. Kawaguchi, A. Lamb, J. Kannala, A. Solin, Y. Bengio, and D. Lopez-Paz, "Interpolation consistency training for semi-supervised learning," *Neural Networks*, vol. 145, pp. 90–106, 2022. DOI: 10.1016/j.neunet.2021.10.008.

[158] Q. Li, Y. Shi, and X. X. Zhu, "Semi-supervised building footprint generation with feature and output consistency training," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–17, 2022. DOI: 10.1109/TGRS.2022.3174636.

[159] L. Yang, L. Qi, L. Feng, W. Zhang, and Y. Shi, "Revisiting weak-to-strong consistency in semi-supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7236–7246.

[160] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "FixMatch: Simplifying semi-supervised learning with consistency and confidence," in *Advances in neural information processing systems 33: Annual conference on neural information processing systems 2020, NeurIPS 2020, december 6-12, 2020, virtual*, 2020.

[161] B. Zhang, Y. Zhang, Y. Li, Y. Wan, H. Guo, Z. Zheng, and K. Yang, "Semi-Supervised deep learning via transformation consistency regularization for remote sensing image semantic segmentation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, pp. 1–15, 2022. DOI: 10.1109/JSTARS.2022.3203750.

[162] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020. DOI: 10.1145/3422622.

[163] W.-C. Hung, Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin, and M.-H. Yang, "Adversarial learning for semi-supervised semantic segmentation," in *British machine vision conference 2018, BMVC 2018, newcastle, UK, september 3-6, 2018*, BMVA Press, 2018, p. 65.

[164] N. Souly, C. Spampinato, and M. Shah, "Semi supervised semantic segmentation using generative adversarial network," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5689–5697. DOI: 10.1109/ICCV.2017.606.

[165] N. A. A. Braham, L. Mou, J. Chanussot, J. Mairal, and X. X. Zhu, "Self supervised learning for few shot hyperspectral image classification," in *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, 2022, pp. 267–270. DOI: 10.1109/IGARSS46834.2022.9884494.

[166] K. Heidler, L. Mou, D. Hu, P. Jin, G. Li, C. Gan, J.-R. Wen, and X. X. Zhu, "Self-supervised audiovisual representation learning for remote sensing data," *International Journal of Applied Earth Observation and Geoinformation*, vol. 116, p. 103 130, 2023. DOI: 10.1016/j.jag.2022.103130.

[167] O. Mañas, A. Lacoste, X. Giró-i-Nieto, D. Vazquez, and P. Rodríguez, "Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data," presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9414–9423.

[168] M. M. i Rabadán, A. Pieropan, H. Azizpour, and A. Maki, "Dense FixMatch: A simple semi-supervised learning method for pixel-wise prediction tasks," 2022. DOI: 10.48550/arXiv.2210.09919.

[169] M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 9630–9640. DOI: 10.1109/ICCV48922.2021.00951.

[170] E. Loebel, M. Scheinert, M. Horwath, A. Humbert, J. Sohn, K. Heidler, C. Liebezeit, and X. X. Zhu, "Calving front monitoring at sub-seasonal resolution: A deep learning application to Greenland glaciers," *The Cryosphere Discussions*, pp. 1–21, 2023. DOI: 10.5194/tc-2023-52.

[171] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-k. Wong, and W.-c. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[172] S. V. Kokelj, T. C. Lantz, J. Kanigan, S. L. Smith, and R. Coutts, "Origin and polycyclic behaviour of tundra thaw slumps, Mackenzie Delta region, Northwest Territories, Canada," *Permafrost and Periglacial Processes*, vol. 20, no. 2, pp. 173–184, 2009. DOI: 10.1002/ppp.642.

[173] S. Oprea, P. Martinez-Gonzalez, A. Garcia-Garcia, J. A. Castro-Vargas, S. Orts-Escolano, J. Garcia-Rodriguez, and A. Argyros, "A review on deep learning techniques for video prediction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 2806–2826, 2022. DOI: 10.1109/TPAMI.2020.3045007.

[174] K. Bi, L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian, "Accurate medium-range global weather forecasting with 3D neural networks," *Nature*, vol. 619, no. 7970, pp. 533–538, 2023. DOI: 10.1038/s41586-023-06185-3.

[175] F. Maussion, A. Butenko, N. Champollion, *et al.*, "The Open Global Glacier Model (OGGM) v1.1," *Geoscientific Model Development*, vol. 12, no. 3, pp. 909–931, 2019. DOI: 10.5194/gmd-12-909-2019.

[176] S. Westermann, T. Ingeman-Nielsen, J. Scheer, *et al.*, "The CryoGrid community model (version 1.0) – a multi-physics toolbox for climate-driven simulations in the terrestrial cryosphere," *Geoscientific Model Development*, vol. 16, no. 9, pp. 2607–2647, 2023. DOI: 10.5194/gmd-16-2607-2023.

[177] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *Journal of Computational Physics*, vol. 378, pp. 686–707, 2019. DOI: 10.1016/j.jcp.2018.10.045.

[178] J. Bolibar, F. Sapienza, F. Maussion, R. Lguensat, B. Wouters, and F. Pérez, "Universal differential equations for glacier ice flow modelling," *Geoscientific Model Development*, vol. 16, no. 22, pp. 6671–6687, 2023. DOI: 10.5194/gmd-16-6671-2023.

[179] J. Kaddour, A. Lynch, Q. Liu, M. J. Kusner, and R. Silva, "Causal Machine Learning: A Survey and Open Problems," 2022. DOI: 10.48550/arXiv.2206.15475.

[180] K. Heidler, L. Mou, and X. X. Zhu, "Seeing the bigger picture: Enabling large context windows in neural networks by combining multiple zoom levels," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, 2021, pp. 3033–3036. DOI: 10.1109/IGARSS47720.2021.9554434.

[181] K. Heidler, L. Mou, E. Loebel, M. Scheinert, S. Lefèvre, and X. X. Zhu, "Deep active contour models for delineating glacier calving fronts," in *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, 2022, pp. 4490–4493. DOI: 10.1109/IGARSS46834.2022.9884819.

A Publications

A.1 HED-UNet: Combined Segmentation and Edge Detection for Monitoring the Antarctic Coastline

Reference

K. Heidler, L. Mou, C. Baumhoer, A. Dietz, and X. X. Zhu, "HED-UNet: Combined segmentation and edge detection for monitoring the antarctic coastline," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022. DOI: 10.1109/TGRS.2021.3064606

Copyright

HED-UNet: Combined Segmentation and Edge Detection for Monitoring the Antarctic Coastline

Konrad Heidler, *Student Member, IEEE*, Lichao Mou, Celia Baumhoer, Andreas Dietz, and Xiao Xiang Zhu, *Fellow, IEEE*

***Abstract*— Deep learning-based coastline detection algorithms have begun to outshine traditional statistical methods in recent years. However, they are usually trained only as single-purpose models to either segment land and water or delineate the coastline. In contrast to this, a human annotator will usually keep a mental map of both segmentation and delineation when performing manual coastline detection. To take into account this task duality, we, therefore, devise a new model to unite these two approaches in a deep learning model. By taking inspiration from the main building blocks of a semantic segmentation framework (UNet) and an edge detection framework (HED), both tasks are combined in a natural way. Training is made efficient by employing deep supervision on side predictions at multiple resolutions. Finally, a hierarchical attention mechanism is introduced to adaptively merge these multiscale predictions into the final model output. The advantages of this approach over other traditional and deep learning-based methods for coastline detection are demonstrated on a data set of Sentinel-1 imagery covering parts of the Antarctic coast, where coastline detection is notoriously difficult. An implementation of our method is available at https://github.com/khdlr/HED-UNet.**

***Index Terms*— Antarctica, edge detection, glacier front, semantic segmentation.**

I. Introduction

CONTRARY to many other landmasses, Antarctica's coastline is fringed by the dynamic glacier and ice shelf fronts continuously changing the coastline location by iceberg calving, which is influenced by both seasonal variations and global climate change. Tracking the advance and retreat of glacier and ice shelf fronts is an important factor for a better understanding of glaciological processes. Furthermore, it is essential to monitor the calving front retreat as it enhances the sea-level contribution of the Antarctic ice sheet due to decreased buttressing effects.

Overall, the length of the Antarctic coastline amounts to around 40 000 km [1], which renders manual delineation infeasible. Especially when observing the developments over multiple time steps for continuous tracking, an automated coastline extraction technique is needed. The recent advances in algorithms and sensing platforms open up new possibilities for the analysis of satellite imagery over large regions, which can be observed in fields as diverse as land cover mapping [2]–[4], bathymetry [5]–[7], urban applications [8]–[12], change detection [13]–[17], and cryosphere research [18]–[22].

This kind of fine-grained analysis is possible because of the availability of satellite imagery with revisit times in the order of days. Regarding data sources, both optical and synthetic aperture radar (SAR) sensors produce imagery suitable for the delineation of the Antarctic coastline [23]. The use of optical imagery in the Antarctic comes with some major drawbacks. Apart from the usual problems with cloud cover, vision is further impeded by polar night and sensor saturation due to the high albedo of ice. To create continuous and gapless observations, data from the Sentinel-1 mission was chosen as the main imagery source. SAR data have often been found to be helpful with the analysis of the cryosphere [24]–[33]. In our case, it allows for near-real-time analysis at a high temporal resolution.

Using SAR data for the task of coastline extraction also imposes some challenges. The speckle present in SAR images makes it harder to pinpoint the exact boundary between land and sea. Furthermore, the backscatter characteristics of glacial ice vary throughout the year, making it hard to distinguish between, e.g. open sea and the higher ice sheet. Therefore, a good model needs to pay additional attention to contextual clues and cannot rely on local information only.

Existing studies for delineating coastlines in general, as well as the Antarctic one, often focus their predictions on either the area of land and sea (*sea–land segmentation*) or the coastline itself (*coastline detection*). However, to the human eye, the two concepts of "area" and "edge" are closely intertwined, making it hard to imagine one without the other. When conducting

Manuscript received September 23, 2020; revised January 22, 2021 and March 1, 2021; accepted March 2, 2021. Date of publication March 23, 2021; date of current version December 13, 2021. This work was supported in part by the Helmholtz Association through the Helmholtz Information and Data Science Incubator project "Artificial Intelligence for Cold Regions" (*AI-CORE*), in part by the Helmholtz Association's Initiative and Networking Fund through Helmholtz AI under Grant ZT-I-PF-5-01—Local Unit "Munich Unit @Aeronautics, Space and Transport (MASTr)," and in part by the German Federal Ministry of Education and Research (BMBF) in the framework of the International Future AI Lab "AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond" under Grant 01DD20001. *(Corresponding author: Xiao Xiang Zhu.)*

Konrad Heidler, Lichao Mou, and Xiao Xiang Zhu are with the Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), 82234 Weßling, Germany, and also with the Data Science in Earth Observation (SiPEO, formerly Signal Processing in Earth Observation), Technical University of Munich (TUM), 80333 Munich, Germany (e-mail: konrad.heidler@dlr.de; lichao.mou@dlr.de; xiaoxiang.zhu@dlr.de).

Celia Baumhoer and Andreas Dietz are with the German Remote Sensing Data Center (DFD), German Aerospace Center (DLR), 82234 Weßling, Germany (e-mail: celia.baumhoer@dlr.de; andreas.dietz@dlr.de).

Digital Object Identifier 10.1109/TGRS.2021.3064606

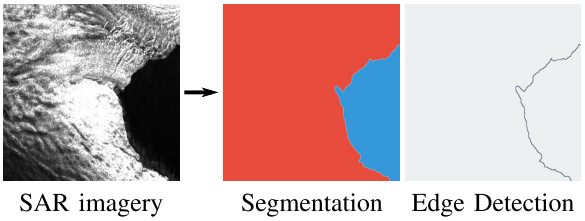


Fig. 1. In coastline detection, the vision tasks of segmentation and edge detection are inseparable.

manual coastline delineation, a human annotator will, therefore, mentally segment the scene into sea and land while searching for the edge between the two at the same time, as shown in Fig. 1.

We hypothesize that taking into account this duality is essential in closing the performance gap between human annotators and automated approaches. In an attempt to more closely model this process, we, thus, introduce a new solution for coastline detection that draws upon the advantages of both segmentation and edge detection approaches. Instead of focusing a predictor on just one of these tasks, our network is trained to jointly perform both tasks at the same time. Inspired by neural architectures for semantic segmentation and edge detection, the model uses an encoder–decoder architecture with skip connections in order to predict segmentation masks and edges at multiple resolutions.

Another observation that we make about coastline detection conducted by humans is the fact that not all areas of a given scene need the same amount of attention to detail. While it is of paramount importance that the coastal regions are precisely mapped, areas further away from the coastline do not receive much attention from a human annotator. By introducing a merging scheme based on hierarchical attention, our model can work in the same way. The intermediate multiresolution predictions are merged using this mechanism to obtain a final output that combines fine-grained low-level outputs with coarser high-level outputs in an efficient way.

Overall, this work's contributions are threefold.

1) Coastline detection is recognized as a dual task. To solve this, a unified theory of segmentation and edge detection is presented. From this, an architecture that implements both semantic segmentation and edge detection is devised.
2) Apart from the narrow coastal strip, there are large regions that require less detailed analysis. This is taken into account by allowing the model to output predictions at different resolution levels. Adding deep supervision for these side outputs improves the training efficiency and generalization performance of the model.
3) In order to dynamically blend between coarse and high-resolution predictions, a hierarchical attention mechanism is used, which takes into account the information available at all levels.

The remainder of this article is organized as follows. Section II gives a brief overview of current methods for coastline detection with a focus on polar regions, as well as existing approaches for combining segmentation and edge detection. Section III presents our proposed HED-UNet architecture. In Section IV, the used data set is introduced. Furthermore, the conducted experiments are explained. Finally, Section V presents numerical results comparing our model to other approaches and ablation studies that analyze the proposed model's elements in detail. Finally, it also includes a discussion of the observed model performance.

II. Related Work

This section will explore the state of the art for coastline detection with a focus on Antarctica. Compared to the general case, the detection of coastlines in the Antarctic requires additional care, as many methods are easily distracted by dynamic sea ice, such as icebergs or ice mélange. Locally, these confounding features can look almost identical to land ice and can, therefore, only be excluded by the additional use of spatial context information.

There are numerous existing approaches for detecting coastlines from satellite imagery. For the biggest part, they can be divided into the aforementioned two classes, differing in the output of interest.

A. Sea–Land Segmentation

In the field of computer vision, semantic segmentation is a central topic. Each pixel is assigned a class, which is to be predicted by the model. This technique is frequently used in remote sensing for various tasks. When the area of either sea or land is of importance, semantic segmentation models are used to distinguish between sea pixels and land pixels.

1) Statistical Methods: In quite a few studies, this has been done by means of statistical analysis. For the Antarctic, the use of a bimodal Gaussian mixture model was proposed, for which parameters are estimated in order to derive an adaptive thresholding scheme. This approach can be applied to both SAR and optical imagery [1]. Similar dynamic thresholding schemes have been applied to different sensors [34]. While easy to implement and fast to evaluate, these methods completely discard the spatial relationships of the pixels, which renders them unfit to deal with the aforementioned issues.

Another localized way of segmenting images that have been applied to sea–land segmentation is given by the watershed algorithm [35]. It treats the pixel intensities as height values and then simulates the resulting surface being flooded with water. Finally, unsupervised clustering methods are helpful in the analysis of complex coastlines [36]. These methods have the benefit of being unsupervised, i.e., requiring no training prior to the evaluation, but the lack of supervision also means that the models cannot be taught to, e.g., ignore icebergs.

2) Deep Learning Methods: With the rise of deep learning in remote sensing [37], convolutional neural networks (CNNs) have been shown to provide superior performance for many tasks, including the one of sea–land segmentation [38]–[40]. Deep convolutional architectures, such as SegNet [41] or UNet [42], leverage contextual information through their encoder–decoder architectures. Thus, as they have more context to base their decisions on, they have the potential

to produce more accurate results than pixelwise or shallow texture-based classifiers. This is of great interest to Antarctic coastline detection due to the aforementioned issues. Current developments in computer vision show a trend toward more complex models for semantic segmentation, which incorporates global information [43] or shape information [44].

Generally, these models require large amounts of labeled data and take quite some time to train. However, they can outperform the previously mentioned methods.

B. Coastline Detection

A closely related task is approached in coastline detection. Instead of segmenting a scene into sea and land, the coastline itself is of primary interest.

1) Edge Tracing: One class of edge detection methods marks the boundaries in the image step by step. After some filtering to highlight the edges, which can be done, e.g., using the Roberts operator [45] or the Sobel operator [46], pixels that are likely to lie on the edge are connected to form the entire boundary. Regarding coastline detection, this approach has been shown to work for SAR data when applying preprocessing steps to account for the nature of the imagery [47]. They can also be connected using a shortest-path algorithm [29] or ridge tracing [48]. Yet, another approach comes from exploiting detection duality. By the nature of the relationship between sea, land, and the coastline, the coastline can be derived from a sea–land segmentation by tracing the transitions between the sea and land classes [49].

While relatively simple, these methods often have some issues regarding robustness. When the tracing procedure takes a wrong turn, it is hard for the algorithm to return to the true boundary.

2) Contour Methods: Active contours, sometimes, also called Snakes [50], are quite similar to the edge tracing approach. Instead of the pixel-by-pixel approach, this class of methods uses an initial curve that is iteratively deformed to minimize an energy function. By choosing the right energy function, this framework can be used to delineate coastlines. For SAR imagery, active contours are able to find coastlines when given a good initialization [51], [52]. These models are sensitive to the provided initialization, meaning that they can converge to local minima that do not represent the desired edge.

3) Level Set Methods: Instead of working with an explicit parameterization of the curve, these methods work with an implicit representation given by a scalar field, in which the zero set represents the boundary [53], [54]. Adaptations of this method for SAR coastline detection use multiple level set iterations to go from coarse to fine delineations [55] or sophisticated preprocessing steps [56] to make the method work for this particular type of imagery.

4) Deep Learning Methods: Only recently, approaches based on deep learning have begun to outperform handcrafted edge detection algorithms. Specialized architectures leverage the framework of CNNs to derive features that predict the presence of edges [57]–[59]. Notably, the previously mentioned Roberts and Sobel operators can be viewed as shallow CNNs with just one layer and a convolutional filter size of 2 and 3, respectively. Therefore, it is only natural that deeper CNNs with more layers are able to outperform these hardcoded edge detection operators.

C. Combining Semantic Segmentation and Edge Detection

A common problem with semantic segmentation models is the blurriness near class boundaries. This likely stems from the fact that the edges make up a minority of the pixels and are, therefore, not well enough represented by the standard pixelwise cross-entropy loss. Thus, the idea of augmenting semantic segmentation approaches with edge information is not a new one.

One way of making a segmentation model aware of edges in the image is by adding an auxiliary loss term that encourages the prediction of crisp edges. This has been shown to work for sea–land segmentation [60].

Surprisingly, simply adding the edge detection task as an auxiliary output for a segmentation model can improve the segmentation results in quite a bit, even without further changes to the model [61]. This approach can also improve sea–land segmentation results in harbor areas [62].

To further improve blurry segmentations, edge masks can be used as the basis for a spatial propagation of class labels. In [63], a segmentation map is initialized using a segmentation network, and at the same time, edges are predicted. These edge masks are then used as the basis for recursive multidirectional label propagation.

For aerial scene classification, the use of an edge detection subnetwork before doing the segmentation has been shown to be beneficial. The detected edge masks are then used as additional input features for the segmentation model. This approach improves the shape accuracy of the resulting segmentation [64].

Contrary to these approaches, we develop a unified theory of segmentation and edge detection. We then identify the components that successful neural networks use to solve either one of these tasks and, finally, devise a model that incorporates the tools necessary to solve both tasks at the same time. The underlying assumption is that both segmentation and edge detection are of equivalent importance for detecting coastlines in satellite imagery.

III. Proposed Method

Implementing the sea–land segmentation task via a UNet segmentation model [42] has become a popular approach for the automatic delineation of coastlines [38]–[40]. Also, in our data set, this method yields good results on the majority of the evaluated scenes [31]. However, oftentimes, the predictions become inaccurate and blurry in areas close to the coastline. As the precise location of the coastline is the central object of our study.

On the other hand, edge detection models excel at delineating the edges in the given images. However, an edge delineation has no concept of “inside” and “outside” by itself, so this output alone is insufficient for labeling sea and land. Furthermore, edge detection models are easily fooled by inland structures of similar appearance to the coastline, as well as

icebergs near the coast. This implies the need for extensive postprocessing and manual corrections.

To put our aforementioned hypotheses into practice, we now introduce a hybrid model for simultaneous prediction of the sea–land segmentation and edge detection of the coastline. Following our observation that humans will usually take into account both the edge information and the textural shape information, we, therefore, propose a combined framework that draws upon the advantages of both these approaches. It takes inspiration from both UNet [42] and HED [57], as well as related architectures by combining key ideas in a very natural way. Therefore, we call our model HED-UNet.

A. Unifying Segmentation and Edge Detection

Regarding the deep learning formulation of the tasks, both segmentation and edge detection are in their nature *dense prediction tasks*, i.e., for each input pixel, an output label needs to be predicted. In the case of segmentation, this is the class label, such as "sea" or "land." For edge detection, it is a classification into the two classes "edge" and "no edge."

This means that, in principle, a segmentation model can be trained to perform edge detection, and vice versa. However, these models were designed for their respective tasks only, meaning that the performance will be degraded when applying them to a different task. In order to construct a model that works well for both tasks, we will, therefore, identify the components of successful architectures for both tasks and find a way to incorporate them into a single multitask model.

1) Segmentation Building Blocks: Some successful semantic segmentation architectures employ the combination of an *encoder* and a *decoder* [41], [42]. The encoder conducts a series of downsampling steps to allow for the aggregation of contextual information at a lower resolution. In turn, the decoder then distributes this information to the individual pixels through a series of upsampling steps.

In a more recent branch of semantic segmentation approaches, the network architecture is divided into a *backbone* network, which calculates feature maps, and one or multiple *prediction heads*, which conduct the final classification based on these feature maps [43], [44], [65].

The contextual aggregation capabilities of an encoder–decoder framework are needed for this task, as some regions can only be classified correctly by the use of contextual clues. At the same time, the backbone-head approach makes it easy to build models that tackle multiple tasks. These considerations lead to the idea of implementing a backbone network that follows the encoder–decoder structure. This has been pioneered for the task of object detection in the framework of *feature pyramid networks* [66]. For our network, we will employ two task-specific prediction heads after calculating a feature pyramid through an encoder–decoder approach.

2) Edge Detection Building Blocks: On the other hand, edge detection frameworks are optimized to provide sharp edge delineations while, at the same time, keeping down the number of false positives. This means that they need to combine the crisp edges predicted at a high resolution with more robust, lower resolution features to reject false positives from the former. Edge detection methods, therefore, often try to strike a balance between predictions or feature maps at different resolutions, which can be done with an architecture that employs an encoder followed by a merging block [57]–[59]. The encoder part is similar to the encoders used in semantic segmentation models; it aggregates contextual information by downsampling. The merging part, however, is a new block that combines the information from different resolution levels after they have been upsampled to the full resolution.

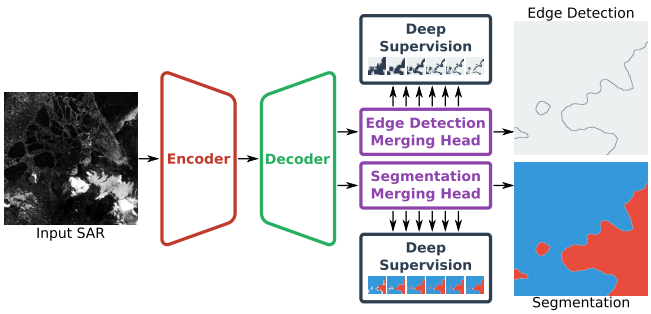


Fig. 2. High-level structure of the proposed framework. First, the encoder and the decoder calculate a pyramid of feature maps. Then, the task-specific merging heads combine this information using the hierarchical attention mechanism.

Looking back at the proposed feature pyramid backbone, such a merging part fulfills the function of a prediction head. This observation leads to the high-level network architecture, as shown in Fig. 2. It is structured in such a way that it contains the components for both segmentation and an edge detection network. After this general structure of the network has been fixed, the detailed layout for each one of these blocks will be outlined in Section III-B.

3) Loss Function: In edge detection, the classes "edge" and "no edge" are highly imbalanced. Therefore, we use an adaptively balancing modification of the binary cross-entropy loss, as proposed in [57]. For a single image with a ground-truth partition into positive pixels Y_+ and negative pixels Y_- and a prediction $\hat{p}$, it is given as

$$\mathcal{L}(\hat{p}) = -\frac{|Y_-| \sum_{j \in Y_+} \log \hat{p}_j}{|Y_+ \cup Y_-|} - \frac{|Y_+| \sum_{j \in Y_-} \log(1 - \hat{p}_j)}{|Y_+ \cup Y_-|}. \quad (1)$$

This loss function gives equal weight to the positive and negative classes, no matter the ratio between the two class sizes. Due to this property, it is fit not only for edge detection but also for semantic segmentation as well. Therefore, it is used as the loss function for both tasks.

B. Architecture Details

Regarding the model details, we start with the encoder–decoder backbone. Conjecturing that the model needs a large spatial context window to base its decisions on, we use a feature pyramid with six resolution levels, corresponding to five downsampling and upsampling steps. In this pyramid, the finest feature map is at the full image

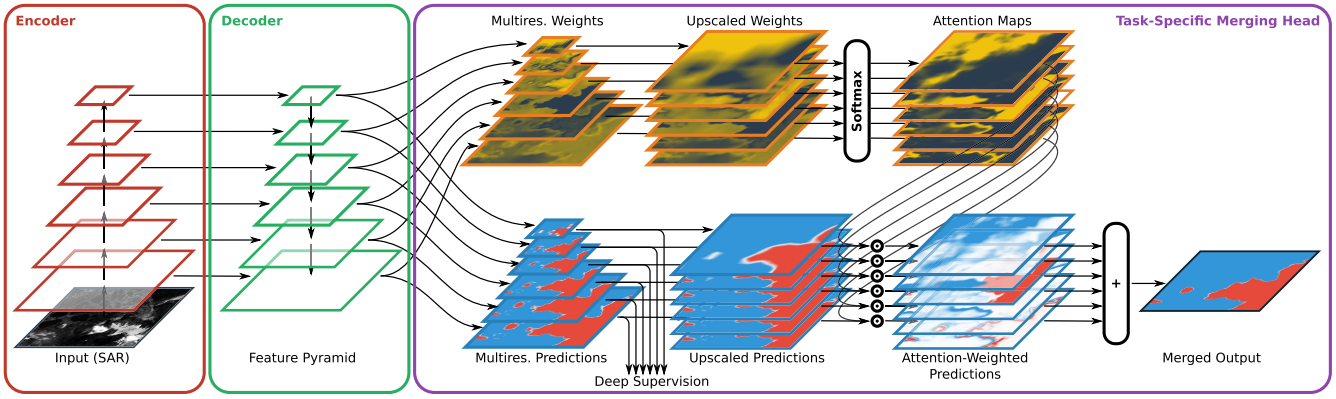


Fig. 3. Architectural details of the proposed network. The full model contains two task-specific merging heads; for clarity, only the segmentation head is shown here. The edge detection head follows the same structure.

resolution, and the coarsest one is at 1/32 resolution. The number 6 was chosen to cover large enough receptive fields needed for the task. Deepening the network even further would lead to receptive fields that exceed the image tiles' extents and did not bring further improvements in our experiments. In the decoder part, the data flows are merged by elementwise addition.

Inspired by the hierarchical nature of the HED architecture [57], we adopt the scheme of predicting coarse representations of the output from within deeper layers. A side output for both segmentation and edge detection is added for each feature map, for a total of six outputs. These multiscale outputs are used in two different ways.

1) Deep Supervision: When building a deep feature pyramid like here, there might not be much motivation for the model to encode meaningful and informative features to the deep, lowest resolution feature maps. In order to explicitly provide this motivation, we train the model to be able to predict the ground truth from each single feature map in the pyramid.

This so-called *deep supervision* [67] is known to improve the learning effectiveness of a neural network, as well as its generalization capabilities. This is achieved by training intermediate network outputs on the ground-truth data to provide additional and more direct training feedback to the earlier layers. In our case, an accordingly downsampled version of the ground-truth segmentation is created for each one of the multiresolution predictions, and the corresponding edges are calculated. Then, these multiscale ground truths are compared with the predictions to provide additional loss terms. The resulting deep supervision encourages the network to better capture larger structures and make use of the available receptive field by encoding meaningful features in the deep layers.

2) Multiscale Fusion: In the next step, these side outputs become part of the merging heads that combine the intermediate outputs into one full-resolution prediction. This is a central point in the original HED architecture [57], so we also implement it in the combined HED-UNet model. In this way, the model has a way of combining fine-grained delineations near the edges with the more robust high-level predictions further away from the edge. The way of merging used in HED is to combine the intermediate predictions using learned weights. However, to further improve the merging performance, we propose the following attention-based merging mechanism.

C. Hierarchical Attention Merging Heads

The final element of the network architecture is the merging heads. In the edge detection frameworks introduced earlier [57]–[59], this is done by featurewise concatenation, followed by a 1×1 convolution to merge the information from different levels. However, in different areas, different fusion behaviors might be needed. In coastal areas, the model might want to use predictions of the highest possible resolution in order to accurately delineate the coastline. However, farther away from the coast, the lower resolution levels can provide a more general assessment of the scene and, thus, lead to better classifications in these areas.

To allow for this adaptive fusion of the multiscale predictions that take into account the confidence at the different granularities, we, therefore, introduce a new fusion procedure based on *attention*. This technique was initially explored in natural language processing as sequential attention among words and tokens [68] and later also applied in computer vision as spatial attention within an image [69].

Inspired by these works, we apply attention to merging multiscale predictions. Here, this mechanism allows the network to focus on the features that it deems most useful for each pixel of the current scene, instead of having fixed weights for feature fusion. Thus, instead of sequential or spatial attention, our attention block allows the model to *attend to different resolution levels*. It works as follows.

For each prediction level, a weight map is created. The weight maps are then upsampled to match the output resolution and turned into a categorical probability map by applying the softmax function over the concatenated resolution levels. To obtain the final prediction, the dot product between the predictions and the attention mask is calculated. This process is visualized in Fig. 3.

For a pyramid of feature maps F_k, the final prediction $\hat{p}$ is, thus, calculated as

$$\hat{p} = \sum_k u\big(f_k(F_k)\big) \cdot \text{softmax}_k\Big(u\big(g_k(F_k)\big)\Big) \tag{2}$$

where u(·) denotes bilinear upsampling to the full output resolution. The functions f_k and g_k denote the multilevel prediction layers and the attention layers, respectively; both are implemented as simple 1×1 convolutional layers.

This approach can be interpreted probabilistically as follows. The intermediate predictions $f_k(F_k)$ can be considered to be maps of Bernoulli probabilities for the output classification at different resolutions. Through the prediction process, these probabilities are conditioned on the input imagery. The original merging procedure with fixed weights corresponds to a mixture model of these Bernoulli maps where the mixture coefficients w_k are learned and fixed. For an input scene X, the predicted probabilities Y are, thus, approximated as

$$P(Y_{ij} \mid X) \approx \sum_k w_k \, P(Y_{ij} \mid X, \text{resolution} = k). \tag{3}$$

Contrary to that, the attention merging corresponds to a mixture model where the mixture coefficients w_{kij} are learned to dynamically depend on the input as well, resulting in the slightly different approximation

$$P(Y_{ij} \mid X) \approx \sum_k w_{kij}(X) \, P(Y_{ij} \mid X, \text{resolution} = k). \tag{4}$$

Notationwise, this might seem like a small change. However, it leads to more flexibility in the resulting probabilistic model, which implies the potential for better classifications.

From the probabilistic perspective, the model training corresponds to a simultaneous maximization of both the side outputs' likelihood and the likelihood of the full mixture under the observed data.

IV. Data Set and Experimental Setup

In order to validate the effectiveness of the suggested improvements, we trained and validated several competing methods and the proposed model on a data set of the Antarctic coast.

A. Data Set

Our data set consists of 16 cropped Sentinel-1 GRD scenes of Antarctica's coastline taken between June 2017 and December 2018 in the sensor's Extra Wide Swath acquisition mode. The spatial distribution of these tiles can be seen in Fig. 4. The data have a resolution of 40 m and dual polarization with HH and HV channels. The cropped scenes have an average size of 7870 × 6572 pixels (315 km × 263 km) and a combined area of around 730 000 km^2. All imageries are processed in the Antarctic Polar Stereographic projection (EPSG:3031) and converted to a decibel. On these scenes, the coastline was manually annotated by experts in order to provide a ground-truth sea–land segmentation and coastline delineation.

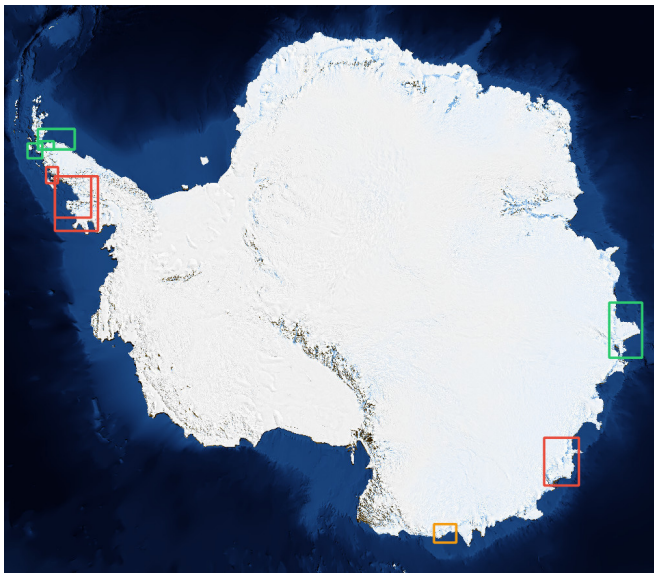


Fig. 4. Spatial distribution of the scenes in the data set. Scenes marked in green were used for model training; scenes marked in red were used for validation purposes. The red area in the top left is the "Antarctic Peninsula" validation site, while the bottom right red area is the "Wilkes Land" validation site. For most locations, data from 2 or 3 different sensing dates were used to allow for an assessment of each model's temporal stability. Marked in yellow is the footprint of the visualization tile in Fig. 7.

The scenes within the data set are clustered in five areas, out of which two were selected as validation areas and completely left them out of the training procedure. This leads to a split of 11 training scenes and five validation scenes. The scenes were all tiled into sections of 768 × 768 pixels with 50% overlap between adjacent tiles to form the training and validation data sets, respectively. In order to improve generalization performance, we employed eightfold data augmentation on the training set. This augmentation technique processes a single tile into the eight different versions that can be obtained by horizontal or vertical mirroring, as well as rotating by multiples of 90°.

B. Evaluated Models

As competitors to our model, we evaluate the following models to provide a baseline.

1) Traditional Methods:

a) Gaussian mixture: The sea–land segmentation method presented in [1] applies dynamic thresholding based on a bimodal mixture of Gaussians.

b) K-medians clustering: An unsupervised sea–land segmentation method presented in [36] employs k-medians clustering of the pixels in a scene on multiple scales.

c) Sobel edges: The coastline detection method presented in [47] applies the Sobel filter, a spatial dilution process, and then a Roberts edge filter.

d) Active contours: Active contours approach for coastline detection based on the Chan-Vese model [54].

2) Deep Learning:

a) HED: The edge detection model from [57].

b) UNet: The segmentation model presented in [42] is known to work well for coastline detection [31].

TABLE I
NUMERICAL RESULTS FOR THE EVALUATED MODELS

Site	Wilkes Land					Antarctic Peninsula				
Metric	Accuracy	mIoU	Deviation	F_1 ODS	F_1 OIS	Accuracy	mIoU	Deviation	F_1 ODS	F_1 OIS
Gaussian Mixture [1]	77.4	63.0	773			74.7	58.8	765		
K-Medians Clustering [36]	55.9	28.0	637			60.5	40.1	560		
Sobel Edges [47]			507	29.0	31.8			644	21.1	20.8
Active Contours [54]			672	21.9	23.5			698	14.6	15.1
HED [57]			341 ± 22	38.4 ± 1.7	41.0 ± 1.0			398 ± 27	28.5 ± 0.8	29.6 ± 0.7
UNet [42]	89.2 ± 3.0	80.6 ± 4.7	271 ± 14			79.3 ± 2.8	65.0 ± 4.4	483 ± 40		
DeepUNet [39]	87.3 ± 6.4	77.6 ± 9.9	287 ± 32			76.9 ± 4.8	61.8 ± 7.5	525 ± 118		
RDUNet [38]	89.2 ± 1.4	80.1 ± 2.2	271 ± 26			78.3 ± 1.2	63.9 ± 2.0	460 ± 73		
HRNet + OCR [43]	89.2 ± 2.5	80.2 ± 4.3	262 ± 35			78.6 ± 1.9	64.6 ± 2.5	467 ± 61		
Gated-SCNN [44]	87.1 ± 0.2	76.8 ± 0.1	297 ± 2	31.6 ± 0.4	34.1 ± 0.1	77.7 ± 1.5	63.0 ± 2.2	471 ± 33	23.0 ± 1.7	25.4 ± 1.7
HED-UNet	92.0 ± 0.8	84.9 ± 1.4	222 ± 23	39.7 ± 1.2	41.6 ± 0.9	80.5 ± 1.6	67.2 ± 2.2	345 ± 24	27.1 ± 1.9	29.4 ± 1.8

c) DeepUNet: A modification of the previous method was developed for sea–land segmentation, as proposed in [39].

d) RDUNet: Another modification of UNet developed for sea–land segmentation, which was proposed in [38].

e) HRNet + OCR: One of the current state-of-the-art models for semantic segmentation in general computer vision [43].

f) Gated-SCNN: It is another recent model for semantic segmentation in general computer vision [44]. This one is particularly interesting, as it also combines segmentation with edge detection.

C. Training Details

The deep learning models were trained on the training data set of Antarctic coastline scenes for 15 epochs on an Nvidia V100 card with 32 GB of video memory. The model weights were optimized by an Adam optimizer using the hyperparameters suggested in [70], namely, a learning rate of 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 10^{-8}$. Due to the large size of the used tiles, the batch size was set to the low number of four samples per batch.

V. RESULTS AND DISCUSSION

The improved performance from our method is quantified using the withheld validation data set. To get informative insights on the actual coastline detection performance, the metrics are calculated only for pixels within 2 km of the true coastline. This way, a distortion of the metrics from noncoastal areas can be avoided.

The two validation areas (Antarctic Peninsula and Wilkes Land; see Fig. 4) are evaluated separately. While the Wilkes Land area can be considered of average difficulty, the Antarctic Peninsula seems to be a very tough location for all of the evaluated models.

For the segmentation approaches, we evaluate the pixelwise accuracy and the mean intersection-over-union metric for the classes of water and land. For edge detection, we calculate the edge F_1 scores at optimal image scale (OIS) and optimal data set scale (ODS). Finally, we calculate an approximate deviation by averaging the distance to the ground-truth coastline over all predicted coastline pixels ("Deviation"). Table I shows the numerical results obtained. The average distance metric can be considered the most important one for this task, as it estimates the overall error between the actual coastline and the predicted coastline. Regarding segmentation performance, the mIoU metric can be considered the primary metric. In order to get a visual impression of some of the models' performance, Fig. 5 shows predictions for a selection of validation tiles. The shown examples are ordered from what we consider easy to hard samples for the models and showcase some of the difficulties with the data set, such as sea ice and confounding backscatter on the higher ice sheet.

A. Model Comparison

First, it is easy to see that the traditional models are not really competitive on this data set. We ascribe this to the repeatedly stated phenomena of icebergs and ice sheet regions with difficult backscatter characteristics. As these models are unsupervised, they simply do not have a way of learning how to deal with such impediments.

Overall, the heterogeneity of the Antarctic coastline is astounding. While the coastline is found pretty well by most models in Wilkes Land, all models have trouble with the scenes from the Antarctic Peninsula.

Among the deep learning-based models, UNet [42] imposes a respectable baseline and even outperforms the more recent models, such as HRNet+OCR [43] and Gated-SCNN [44], in some of the evaluated metrics. Even though the latter also has a side output for edge detection, we find that its edge detection results fall short in comparison to HED [57] and HED-UNet. A reason for this might be the lack of a pretrained backbone network for Sentinel-1 data, which forced us to randomly initialize the backbone and train it alongside the rest of the network. Furthermore, this model was optimized for the segmentation of scenes with many different classes and small objects, which is needed for tasks, such as autonomous driving. In our use case, however, there are only two classes that are nearly equal in area, imposing a very different data distribution.

The ultimate goal of this study is to delineate the coastline as accurately as possible. In the corresponding average deviation metric, the proposed HED-UNet model outshines the alternative approaches, especially in the Antarctic Peninsula validation area. This confirms our assumptions that,

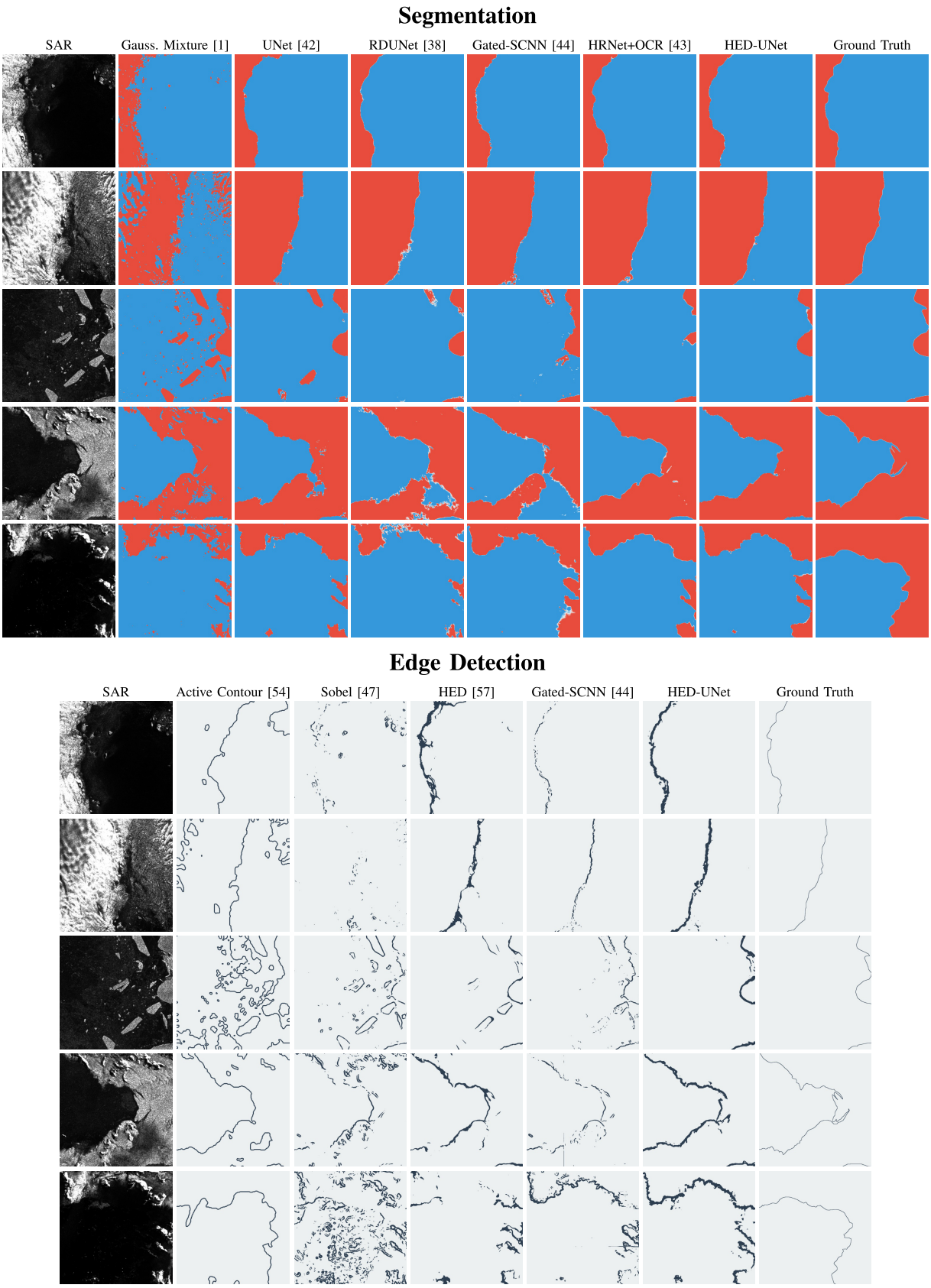


Fig. 5. Qualitative results comparing the evaluated models on unseen validation tiles. In order to provide an informative visualization, the visualized tiles were selected to represent the full spectrum of (Top) easy to (Bottom) hard scenes within the validation set.

TABLE II
NUMERICAL RESULTS FOR THE ABLATIONS

				Wilkes Land					Antarctic Peninsula				
Data	Deep Sup.	Levels	Merging	Accuracy	mIoU	Deviation	F_1 ODS	F_1 OIS	Accuracy	mIoU	Deviation	F_1 ODS	F_1 OIS
SAR	Yes	5	Attention	90.3 ± 1.3	82.2 ± 2.1	239 ± 16	37.2 ± 0.7	38.7 ± 0.9	77.0 ± 1.0	62.5 ± 1.3	379 ± 45	25.4 ± 1.2	27.0 ± 1.0
SAR	No	6	Attention	88.5 ± 1.4	79.2 ± 2.3	954 ± 7	7.1 ± 0.5	7.1 ± 0.5	80.3 ± 1.6	66.8 ± 2.1	895 ± 7	7.4 ± 0.2	7.5 ± 0.2
SAR	Yes	6	None	89.7 ± 1.2	81.1 ± 1.9	284 ± 37	37.5 ± 1.2	39.7 ± 1.3	81.8 ± 2.0	69.0 ± 2.9	378 ± 35	28.0 ± 1.8	30.2 ± 1.8
SAR	Yes	6	Learned	89.9 ± 2.5	81.6 ± 3.9	236 ± 14	37.0 ± 2.2	38.6 ± 2.2	81.4 ± 1.1	68.3 ± 1.6	391 ± 12	26.6 ± 1.4	29.1 ± 1.4
SAR	Yes	6	Attention	92.0 ± 0.8	84.9 ± 1.4	222 ± 23	39.7 ± 1.2	41.6 ± 0.9	80.5 ± 1.6	67.2 ± 2.2	345 ± 24	27.1 ± 1.9	29.4 ± 1.8
SAR+DEM	Yes	6	Attention	92.9 ± 1.4	86.7 ± 2.4	226 ± 47	35.1 ± 3.4	36.0 ± 3.5	91.6 ± 1.6	84.6 ± 2.7	210 ± 9	30.7 ± 2.1	31.4 ± 2.7

for this specific task, our considerations lead to increased performance.

B. Network Depth and Deep Supervision

As a means of quantifying the improvements made to the architecture, we evaluate versions of our model with only some of the improvements applied. The results of this ablation study are displayed in Table II.

For a fair comparison with UNet-based models, we evaluate the performance when only five resolution levels are used instead of six, corresponding to four downsampling and upsampling steps instead of five. While this setup performs slightly worse than the full HED-UNet, it still outperforms the baseline methods.

Regarding deep supervision, we can see that it is of paramount importance for edge detection performance. Without it, the model is barely able to predict the presence of edges. What is more, the coastline is often missed completely due to this poor edge detection performance. On the other hand, deep supervision does not seem to alter the performance of the semantic segmentation task much. This is in line with the original models that we took inspiration from. While the segmentation model UNet [42] does not employ deep supervision, the edge detection model HED [57] makes heavy use of it.

C. Merging Strategies

After adding the deep supervision, we evaluate different merging strategies.

a) None: First, we evaluate a configuration where just the last layer of the decoder is used for the predictions (denoted "None"). This corresponds to the workings of a UNet [42] model with two final prediction layers: one for each task.

b) Learned: Second, we evaluate the performance of the learned merging strategy, as originally proposed in [57]. Here, a prediction is computed for each resolution level in the feature pyramid. These predictions are then upsampled to full resolution and concatenated. After this, a 1×1 convolutional layer with learned weights computes the final prediction from the concatenated prediction stack.

c) Attention: The last strategy is the hierarchical attention merging introduced in Section III-C, which does not rely on fixed weights, such as the previous strategy, but computes the merging weights dynamically for each pixel within each scene.

From our results, learned merging does not improve much over no merging for segmentation and even performs a bit worse for edge detection. The average deviation improves quite a bit in Wilkes Land but worsens a bit on the Antarctic Peninsula in return. We ascribe this to the large differences in the validation areas. As the merging coefficients are fixed for the "Learned" approach, this might hint at the fact that the model learns coefficients that work well for Wilkes Land, but less so for the Antarctic Peninsula.

This issue is overcome by our newly proposed attention merging strategy, which can adapt to the different scenes. It can learn to find good sets of merging coefficients for both Wilkes Land and the Antarctic Peninsula even though the optimal values for each one might be different.

Fig. 6 shows that the model indeed directs its attention in an adaptive fashion as we conjectured. Overall, a mix of all resolution levels is used to compute the final output. On tiles that are completely covered by one of the two classes, the attention shifts a bit toward the lower resolution levels, as they tend to provide more robust predictions. For pixels on the edge, the model heavily focuses on the highest available resolution level, in order to arrive at accurate delineations in these regions.

D. DEM Experiments

Furthermore, we look into including digital elevation data from the TanDEM-X elevation model [71]. We conjecture that this secondary data source can help the model better deject misclassifications from icebergs or dry-snow facies of the higher ice sheet, which have confounding SAR backscatter.

To discourage the model from directly reproducing the coastline implied by the elevation model, we decided to downsample the DEM's resolution to 640 m. This resolution is coarse enough to not make a segmentation based on the DEM alone competitive to the non-DEM models, which has an average deviation of less than 300 m. Furthermore, it allows for easy feature fusion, as it corresponds to the resolution of the feature map at $1/16$ of the full resolution. Therefore, it is simply concatenated to the feature map after the fourth downsampling step in the encoder.

The results when including the DEM are displayed as the last ablation in Table II. On the very hard scenes of the Arctic Peninsula, this additional information helps the model by a large margin, boosting the average deviation from 345 to 210 m. However, the story is different for Wilkes Land. Here, the deviation worsens slightly, and the edge detection metrics go down considerably.

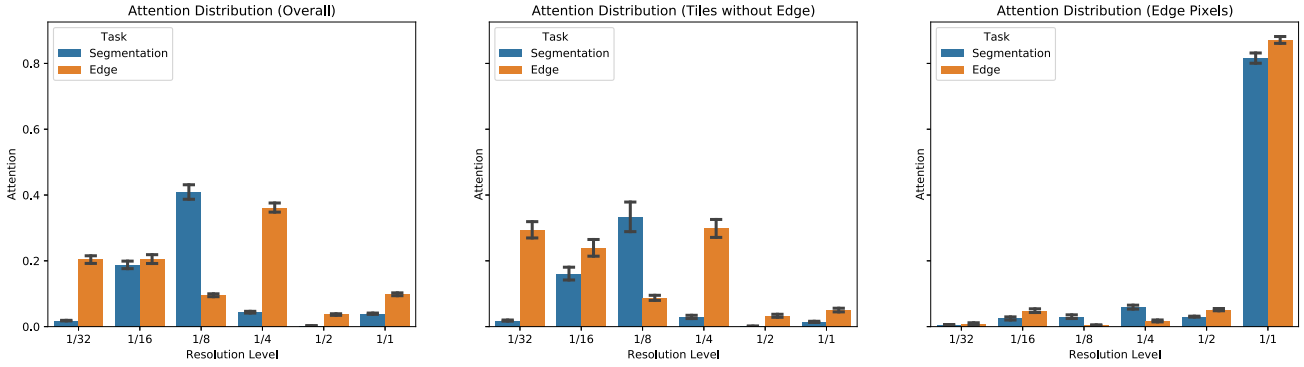


Fig. 6. Amount of attention spent on the different resolution levels. Each plot analyzes a specific class of pixels in the validation data set—from left to right: Average over all pixels, average over pixels from edge-less tiles, and average over all edge pixels.

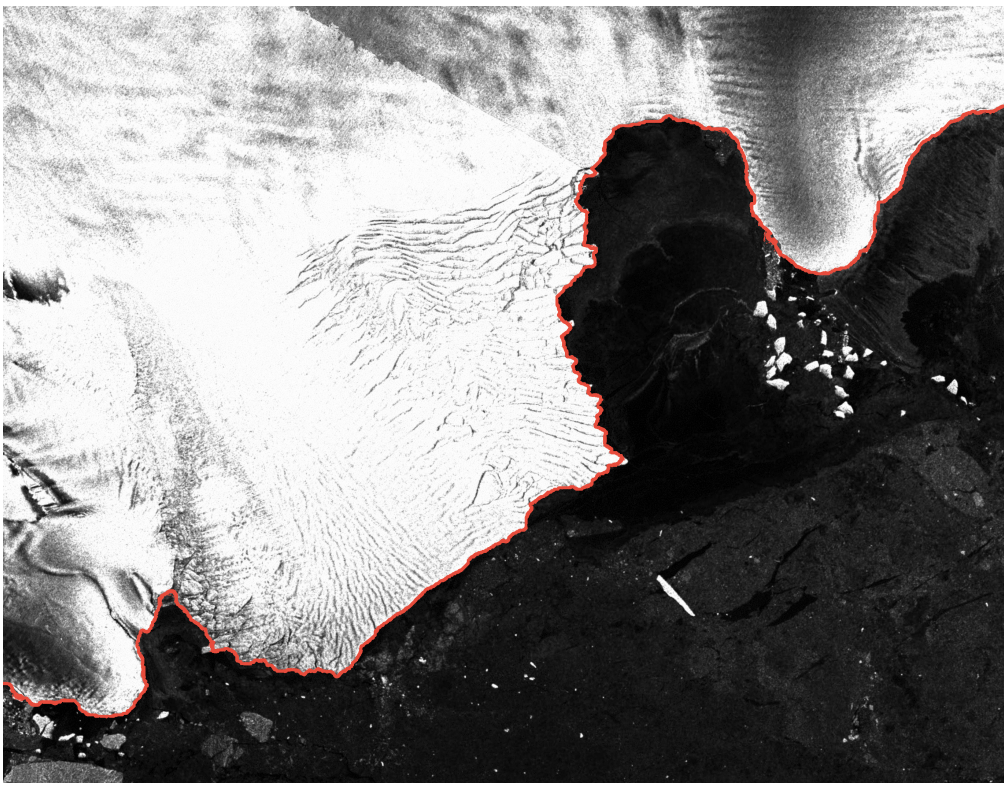


Fig. 7. Section of George V Coast with Cape Hudson in the bottom left, imagery mosaiced from Sentinel-1 takes in early 2019. This scene is both temporally and spatially separated from the training and validation sets used. Overlaid in red is the coastline predicted by the HED-UNet model.

This is a strong indicator that the model is, indeed, overfitting on the DEM to some extent. For example, in some highly dynamic coastal regions, the model will be confused when the DEM and SAR imagery are contradictory.

Thus, all in all, the inclusion of DEM data can be beneficial but needs to be done very carefully to prevent the model from overfitting to the DEM alone.

E. Limitations

Even though the newly proposed model outperforms the baselines on nearly all validation scenes, there are still cases where the results are not perfect. Most misclassifications can be attributed to one of two failure modes, which we will now briefly discuss. Visual examples for these failure modes can be seen in Fig. 8.

1) Sea Ice: The large receptive field and multitask training help alleviate the issue of wrongly classified sea ice. However, very large icebergs and areas of ice mélange can still throw off the proposed model. The first failure example displays such an area where large clusters of sea ice confuse the model.

2) Missing Context: For areas close to the border of a tile, the model sometimes does not have enough contextual information to correctly classify them. This can be observed in the second failure visualization, where a patch of sea ice directly next to the tile border is wrongly classified as land.

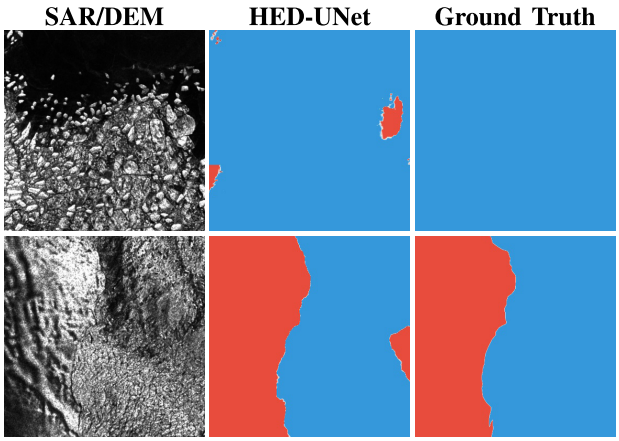


Fig. 8. Failure modes of the proposed model. (Top) Confusion from a very large cluster of sea ice. (Bottom) Confusion due to missing context at the border of the tile.

Overall, these failures do not occur often throughout the data set and apply not only to the HED-UNet models but to the other compared models as well. Especially, the first one requires much human interpretation in a large spatial context, which is difficult for a neural network to achieve without general reasoning capabilities.

F. Effective Receptive Fields

Deep CNNs, such as the ones used in our experiments, have very large theoretical receptive fields. It is conjectured that, while long-range connections are theoretically possible in these networks, networks will often ignore them in favor of short-range connections.

To assess how much of the spatial context is actually used by a CNN, its so-called *effective receptive field* (ERF) can be estimated [72]. This is done by analyzing the expected gradient magnitude of each input pixel with respect to a central output pixel. For a CNN f and a sequence of input images I_k, one, therefore, looks at the values of

$$E = \frac{1}{n}\sum_{k=1}^{n}\left|\nabla_{I_k} f(I_k)_{i,j}\right| \tag{5}$$

for a central output pixel (i, j). If, for an input pixel (x, y), the value $E_{x,y}$ is nonnegligible, then this pixel will influence the output predictions at position (i, j). The spatial distribution of these relevant pixels is then called the ERF.

As the gradient magnitude gives insight on how much the prediction changes in response to a change in the input, the ERF allows for a measurement of the spatial context used by the model. A model with a larger ERF bases its decisions on a larger spatial context than one with a small ERF.

We conjectured that, for the task of Antarctic coastline detection, a model needs to take a large context window into account. Indeed, there seems to be a correlation between a larger ERF and better validation scores for this task.

It can be observed that the UNet model is limited by its theoretical receptive field. Its ERF is forced into an almost quadratic shape because of this. The ERF of the Gated-SCNN model is particularly interesting with its fractal-like shape. We conjecture that this is due to the Atrous Spatial Pyramid Pooling block used in the network architecture, which makes heavy use of dilated convolutions.

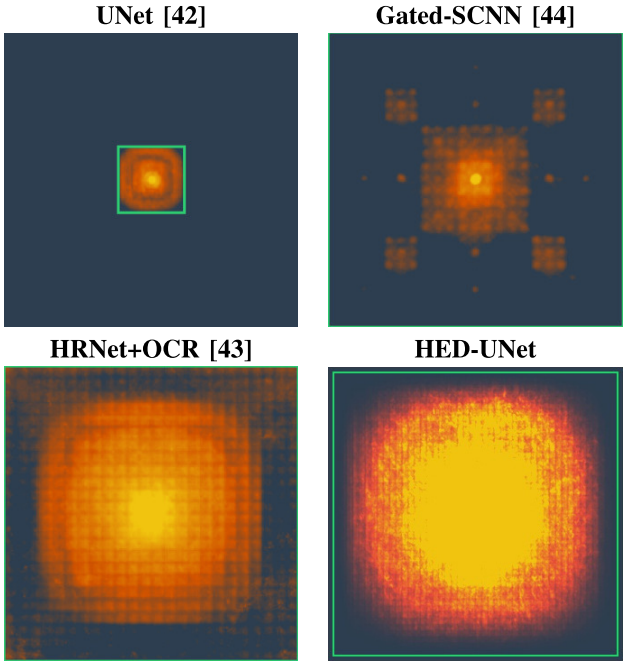


Fig. 9. ERFs of some tested models for the prediction of a central pixel, visualized in image space. Theoretical receptive fields are outlined in green. Note that the theoretical receptive fields of Gated-SCNN and HRNet+OCR are larger than the used patch size of 768×768.

Finally, the HRNet+OCR and HED-UNet models employ a very large ERF, which, once more, supports our assumption that a large receptive field is needed for coastline detection in Antarctica.

VI. Conclusion

In this article, we introduced a model for simultaneous segmentation and edge detection. The proposed HED-UNet learns to exploit the synergies between the two tasks and, thereby, manages to surpass both edge detection and semantic segmentation baselines. By the use of deep supervision, we encourage the model to encode meaningful features in its deep layers, which allows for more general predictions. Finally, the proposed attention merging heads allow for better learning performance and more robust classifications.

Compared to approaching the task with a regular UNet, the presented network architecture only requires a little additional computational cost. Most of the performance gains stem from the adapted training procedure and a few additional layers, which do not require many computational resources compared to the layers already present.

While it is not a general-purpose model, we show that our proposed improvements to the model are, indeed, beneficial for the task of coastline detection. Visual and numerical inspections of the results confirm our assumption that the combination of the two tasks helps the model better grasp the concept of a coastline.

Our model can be applied to coastline detection tasks not only in polar regions but to coastal regions worldwide. Furthermore, we are convinced that the approach taken by HED-UNet will greatly benefit other tasks requiring an edge detection approach in combination with semantic segmentation. Possible applications include the mapping of building footprints, roads, and bodies of water, such as lakes or rivers.

ACKNOWLEDGMENT

The authors thank the European Union Copernicus program for providing Sentinel-1. TanDEM-X elevation data are courtesy of the German Aerospace Center (DLR).

REFERENCES

[1] H. Liu and K. C. Jezek, "A complete high-resolution coastline of antarctica extracted from orthorectified radarsat SAR imagery," *Photogramm. Eng. Remote Sens.*, vol. 70, no. 5, pp. 605–616, May 2004.

[2] J. Geng, H. Wang, J. Fan, and X. Ma, "Deep supervised and contractive neural network for SAR image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 4, pp. 2442–2459, Apr. 2017.

[3] C. Robinson *et al.*, "Large scale high-resolution land cover mapping with multi-resolution data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12726–12735.

[4] X.-Y. Tong *et al.*, "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sens. Environ.*, vol. 237, Feb. 2020, Art. no. 111322.

[5] F. Eugenio, J. Marcello, and J. Martin, "High-resolution maps of bathymetry and benthic habitats in shallow-water environments using multispectral remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 7, pp. 3539–3549, Jul. 2015.

[6] J. Liang, J. Zhang, Y. Ma, and C.-Y. Zhang, "Derivation of bathymetry from high-resolution optical satellite imagery and USV sounding data," *Mar. Geodesy*, vol. 40, no. 6, pp. 466–479, Nov. 2017.

[7] M. Erena, J. A. Domínguez, J. F. Atenza, S. García-Galiano, J. Soria, and Á. Pérez-Ruzafa, "Bathymetry time series using high spatial resolution satellite images," *Water*, vol. 12, no. 2, p. 531, Feb. 2020.

[8] Y. Long, Y. Gong, Z. Xiao, and Q. Liu, "Accurate object localization in remote sensing images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2486–2498, May 2017.

[9] N. Audebert, B. Le Saux, and S. Lefèvre, "Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 140, pp. 20–32, Jun. 2018.

[10] L. Mou and X. X. Zhu, "Vehicle instance segmentation from aerial image and video using a multitask learning residual fully convolutional network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6699–6711, Nov. 2018.

[11] L. Mou, Y. Hua, and X. X. Zhu, "Relation matters: Relational context-aware fully convolutional network for semantic segmentation of high-resolution aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 7557–7569, Nov. 2020.

[12] Q. Li, Y. Shi, X. Huang, and X. X. Zhu, "Building footprint generation by integrating convolution neural network with feature pairwise conditional random field (FPCRF)," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 7502–7519, Nov. 2020.

[13] D. Wen, X. Huang, L. Zhang, and J. A. Benediktsson, "A novel automatic change detection method for urban high-resolution remotely sensed imagery based on multiindex scene representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 1, pp. 609–625, Jan. 2016.

[14] O. Ajadi, F. Meyer, and P. Webley, "Change detection in synthetic aperture radar images using a multiscale-driven approach," *Remote Sens.*, vol. 8, no. 6, p. 482, Jun. 2016.

[15] Z. Y. Lv, T. F. Liu, P. Zhang, J. A. Benediktsson, T. Lei, and X. Zhang, "Novel adaptive histogram trend similarity approach for land cover change detection by using bitemporal very-high-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 9554–9574, Dec. 2019.

[16] B. Du, L. Ru, C. Wu, and L. Zhang, "Unsupervised deep slow feature analysis for change detection in multi-temporal remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 9976–9992, Dec. 2019.

[17] C. Zhang, S. Wei, S. Ji, and M. Lu, "Detecting large-scale urban land cover changes from very high resolution remote sensing images using CNN-based classification," *ISPRS Int. J. Geo-Inf.*, vol. 8, no. 4, p. 189, Apr. 2019.

[18] M. Engram, C. D. Arp, B. M. Jones, O. A. Ajadi, and F. J. Meyer, "Analyzing floating and bedfast lake ice regimes across arctic alaska using 25 years of space-borne SAR imagery," *Remote Sens. Environ.*, vol. 209, pp. 660–676, May 2018.

[19] I. Sasgen, H. Konrad, V. Helm, and K. Grosfeld, "High-resolution mass trends of the antarctic ice sheet through a spectral combination of satellite gravimetry and radar altimetry observations," *Remote Sens.*, vol. 11, no. 2, p. 144, Jan. 2019.

[20] D. O. Dammann, L. E. B. Eriksson, A. R. Mahoney, H. Eicken, and F. J. Meyer, "Mapping pan-arctic landfast sea ice stability using Sentinel-1 interferometry," *Cryosphere*, vol. 13, no. 2, pp. 557–577, Feb. 2019.

[21] I. Nitze, G. Grosse, B. M. Jones, V. E. Romanovsky, and J. Boike, "Remote sensing quantifies widespread abundance of permafrost region disturbances across the arctic and subarctic," *Nature Commun.*, vol. 9, no. 1, pp. 1–11, Dec. 2018.

[22] J. E. Anderson, T. A. Douglas, R. A. Barbato, S. Saari, J. D. Edwards, and R. M. Jones, "Linking vegetation cover and seasonal thaw depths in interior Alaska permafrost terrains using remote sensing," *Remote Sens. Environ.*, vol. 233, Nov. 2019, Art. no. 111363.

[23] C. Baumhoer, A. Dietz, S. Dech, and C. Kuenzer, "Remote sensing of antarctic glacier and ice-shelf front dynamics—A review," *Remote Sens.*, vol. 10, no. 9, p. 1445, Sep. 2018.

[24] T. Strozzi, A. Luckman, T. Murray, U. Wegmuller, and C. L. Werner, "Glacier motion estimation using SAR offset-tracking procedures," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 11, pp. 2384–2391, Nov. 2002.

[25] G. Vasile *et al.*, "High-resolution SAR interferometry: Estimation of local frequencies in the context of alpine glaciers," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 4, pp. 1079–1090, Apr. 2008.

[26] E. Erten, "Glacier velocity estimation by means of a polarimetric similarity measure," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 6, pp. 3319–3327, Jun. 2013.

[27] V. Akbari, A. P. Doulgeris, and T. Eltoft, "Monitoring glacier changes using multitemporal multipolarization SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 6, pp. 3729–3741, Jun. 2014.

[28] S. Lang, X. Liu, B. Zhao, X. Chen, and G. Fang, "Focused synthetic aperture radar processing of ice-sounding data collected over the east antarctic ice sheet via the modified range migration algorithm using curvelets," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4496–4509, Aug. 2015.

[29] L. Krieger and D. Floricioiu, "Automatic glacier calving front delineation on TerraSAR-X and Sentinel-1 SAR imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2017, pp. 2817–2820.

[30] V. Akbari and C. Brekke, "Iceberg detection in open and ice-infested waters using C-band polarimetric synthetic aperture radar," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 1, pp. 407–421, Jan. 2018.

[31] C. A. Baumhoer, A. J. Dietz, C. Kneisel, and C. Kuenzer, "Automated extraction of antarctic glacier and ice shelf fronts from Sentinel-1 imagery using deep learning," *Remote Sens.*, vol. 11, no. 21, p. 2529, Oct. 2019.

[32] E. Zhang, L. Liu, and L. Huang, "Automatically delineating the calving front of Jakobshavn isbræ from multitemporal TerraSAR-X images: A deep learning approach," *Cryosphere*, vol. 13, no. 6, pp. 1729–1741, Jun. 2019.

[33] Y. Mohajerani, M. Wood, I. Velicogna, and E. Rignot, "Detection of glacier calving margins with convolutional neural networks: A case study," *Remote Sens.*, vol. 11, no. 1, p. 74, Jan. 2019.

[34] B. W. J. Miles, C. R. Stokes, and S. S. R. Jamieson, "Simultaneous disintegration of outlet glaciers in porpoise bay (Wilkes Land), east antarctica, driven by sea ice break-up," *Cryosphere*, vol. 11, no. 1, pp. 427–442, Feb. 2017.

[35] Y. Liu *et al.*, "Ocean-driven thinning enhances iceberg calving and retreat of antarctic ice shelves," *Proc. Nat. Acad. Sci. USA*, vol. 112, no. 11, pp. 3263–3268, Mar. 2015.

[36] M. Schmitt, G. Baier, and X. X. Zhu, "Potential of nonlocally filtered pursuit monostatic TanDEM-X data for coastline detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 148, pp. 130–141, Feb. 2019.

[37] X. X. Zhu *et al.*, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[38] P. Shamsolmoali, M. Zareapoor, R. Wang, H. Zhou, and J. Yang, "A novel deep structure U-Net for sea-land segmentation in remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 9, pp. 3219–3232, Sep. 2019.

[39] R. Li *et al.*, "DeepUNet: A deep fully convolutional network for pixel-level sea-land segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 11, pp. 3954–3962, Nov. 2018.

[40] Z. Chu, T. Tian, R. Feng, and L. Wang, "Sea-land segmentation with res-UNet and fully connected CRF," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. IGARSS*, Jul. 2019, pp. 3840–3843.

[41] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[42] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Oct. 2015, pp. 234–241.

[43] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," in *Computer Vision—ECCV*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham, Switzerland: Springer, 2020, pp. 173–190.

[44] T. Takikawa, D. Acuna, V. Jampani, and S. Fidler, "Gated-SCNN: Gated shape CNNs for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5229–5238.

[45] L. G. Roberts, "Machine perception of three-dimensional solids," Ph.D. dissertation, Massachusetts Inst. Technology, Cambridge, MA, USA, 1963.

[46] W. K. Pratt, "Edge detection," in *Digital Image Processing*. Hoboken, NJ, USA: Wiley, 2006, pp. 465–533.

[47] J.-s. Lee and I. Jurkevich, "Coastline detection and tracing in SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 28, no. 4, pp. 662–668, Jul. 1990.

[48] D. Wang and X. Liu, "Coastline extraction from SAR images using robust ridge tracing," *Mar. Geodesy*, vol. 42, no. 3, pp. 286–315, May 2019.

[49] M. Modava, G. Akbarizadeh, and M. Soroosh, "Integration of spectral histogram and level set for coastline detection in SAR images," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 55, no. 2, pp. 810–819, Apr. 2019.

[50] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *Int. J. Comput. Vis.*, vol. 1, no. 4, pp. 321–331, Jan. 1988.

[51] T. Klinger, M. Ziems, C. Heipke, H. W. Schenke, and N. Ott, "Antarctic coastline detection using snakes," *Photogramm. Fernerkund. Geoinf.*, vol. 2011, no. 6, pp. 421–434, Dec. 2011.

[52] C. Liu, Y. Xiao, and J. Yang, "A coastline detection method in polarimetric SAR images mixing the region-based and edge-based active contour models," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3735–3747, Jul. 2017.

[53] S. Osher and J. A. Sethian, "Fronts propagating with curvature-dependent speed: Algorithms based on hamilton-jacobi formulations," *J. Comput. Phys.*, vol. 79, no. 1, pp. 12–49, Nov. 1988.

[54] T. Chan and L. Vese, "An active contour model without edges," in *Scale-Space Theories in Computer Vision*. Berlin, Germany: Springer, Sep. 1999, pp. 141–151.

[55] C. Liu, J. Yang, J. Yin, and W. An, "Coastline detection in SAR images using a hierarchical level set segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 11, pp. 4908–4920, Nov. 2016.

[56] M. Modava and G. Akbarizadeh, "A level set based method for coastline detection of SAR images," in *Proc. 3rd Int. Conf. Pattern Recognit. Image Anal. (IPRIA)*, Apr. 2017, pp. 253–257.

[57] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1395–1403.

[58] Y. Liu, M.-M. Cheng, X. Hu, K. Wang, and X. Bai, "Richer convolutional features for edge detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3000–3009.

[59] X. Soria, E. Riba, and A. Sappa, "Dense extreme inception network: Towards a robust CNN model for edge detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1923–1932.

[60] D. Cheng, G. Meng, G. Cheng, and C. Pan, "SeNet: Structured edge network for sea-land segmentation," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 2, pp. 247–251, Feb. 2017.

[61] Z. Jiang, Z. Chen, K. Ji, and J. Yang, "Semantic segmentation network combined with edge detection for building extraction in remote sensing images," *Proc. SPIE*, vol. 11430, Feb. 2020, Art. no. 114300D.

[62] D. Cheng, G. Meng, S. Xiang, and C. Pan, "FusionNet: Edge aware deep convolutional networks for semantic segmentation of remote sensing harbor images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 12, pp. 5769–5783, Dec. 2017.

[63] L.-C. Chen, J. T. Barron, G. Papandreou, K. Murphy, and A. L. Yuille, "Semantic image segmentation with task-specific edge detection using CNNs and a discriminatively trained domain transform," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4545–4554.

[64] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla, "Classification with an edge: Improving semantic image segmentation with boundary detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 135, pp. 158–172, Jan. 2018.

[65] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*. [Online]. Available: http://arxiv.org/abs/1706.05587

[66] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.

[67] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Proc. Int. Conf. Artif. Intell. Stat. (AISTATS)*, vol. 38, G. Lebanon and S. V. N. Vishwanathan, Eds. San Diego, CA, USA: Proeedings of Machine Learning Research, May 2015, pp. 562–570.

[68] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, p. 11.

[69] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7794–7803.

[70] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent. ICLR*, 2015, pp. 1–15.

[71] P. Rizzoli *et al.*, "Generation and performance assessment of the global TanDEM-X digital elevation model," *ISPRS J. Photogramm. Remote Sens.*, vol. 132, pp. 119–139, Oct. 2017.

[72] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst. (NIPS)*, Dec. 2016, pp. 4905–4913.

Konrad Heidler (Student Member, IEEE) received the bachelor's degree in mathematics and the master's degree in mathematics in data science from the Technical University of Munich (TUM), Munich, Germany, in 2017, and 2020, respectively. He is pursuing the Ph.D. degree with the German Aerospace Center (DLR), Weßling, Germany, and TUM.

His main research interests are remote sensing, computer vision, and mathematical foundations of machine learning. His research work focuses on the application of deep learning in polar regions.

Lichao Mou received the bachelor's degree in automation from the Xi'an University of Posts and Telecommunications, Xi'an, China, in 2012, the master's degree in signal and information processing from the University of Chinese Academy of Sciences (UCAS), Beijing, China, in 2015, and the Dr.Ing. degree from the Technical University of Munich (TUM), Munich, Germany, in 2020.

In 2015, he spent six months at the Computer Vision Group, University of Freiburg, Freiburg im Breisgau, Germany. Since 2019, he has been an AI Consultant for the Helmholtz Artificial Intelligence Cooperation Unit (HAICU), Munich. In 2019, he was a Visiting Researcher with the Cambridge Image Analysis Group (CIA), University of Cambridge, Cambridge, U.K. From 2019 to 2020, he was a Research Scientist with Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Weßling, Germany. He is a Guest Professor with the Munich AI Future Lab AI4EO, TUM, and the Head of Visual Learning and Reasoning Team, Department "EO Data Science," IMF, DLR.

Dr. Mou was a recipient of the First Place at the 2016 IEEE GRSS Data Fusion Contest and a finalist for the Best Student Paper Award at the 2017 Joint Urban Remote Sensing Event and the 2019 Joint Urban Remote Sensing Event.

Celia Baumhoer received the bachelor's degree in physical geography from the University of Erlangen-Nuremberg, Erlangen, Germany, in 2014, and the master's degree in geography (environmental systems in transition) from the Rheinische Friedrich-Wilhelms-University Bonn, Bonn, Germany, in 2017. She is pursuing the Ph.D. degree with the German Aerospace Center (DLR), Weßling, Germany, and the Julius-Maximilians-University Würzburg, Würzburg, Germany.

Since 2020, she has been a Scientific Assistant with the Group "Polar and Cold Regions," German Remote Sensing Data Center (DFD), DLR, working on AI-applications for cold regions. Her research interests include synthetic aperture radar (SAR) remote sensing and machine learning with a special focus on the cryosphere and Antarctic glaciers in particular.

Dr. Baumhoer was a recipient of the DAAD Rise Scholarship for a research stay in Lafayette, LA, USA, and the ASTO-Förderpreis for excellent Ph.D. students.

Andreas Dietz received the diploma and Dr.rer.nat. degrees from Julius-Maximilians-University Würzburg, Würzburg, Germany, in 2009 and 2013, respectively.

He has been the Head of the Group "Polar and Cold Regions," Department "Land Surface Dynamics," German Remote Sensing Data Center (DFD), German Aerospace Center (DLR), Weßling, Germany, since 2018; this group focuses on the development of methods to quantify the impact of climate change on the cryosphere based on remote sensing data. Through this research, the DLR Global SnowPack has been developed, which is an operational, globally available daily snow cover product. His research interests include remote sensing, Earth observation, and climate change with a focus on the cryosphere.

Dr. Dietz was a recipient of the Eastern Snow Conference Wiesnet Medal in 2013 and the Helmut Rott Award in 2015.

Xiao Xiang Zhu (Fellow, IEEE) received the M.Sc., Dr.Ing., and Habilitation degrees in signal processing from the Technical University of Munich (TUM), Munich, Germany, in 2008, 2011, and 2013, respectively.

She was a Guest Scientist or a Visiting Professor with the Italian National Research Council (CNR-IREA), Naples, Italy, Fudan University, Shanghai, China, The University of Tokyo, Tokyo, Japan, and the University of California at Los Angeles, Los Angeles, CA, USA, in 2009, 2014, 2015, and 2016, respectively. She is the Professor of data science in Earth observation (former: signal processing in Earth observation) with TUM and the Head of the Department "EO Data Science," Remote Sensing Technology Institute, German Aerospace Center (DLR), Weßling, Germany. Since 2019, she has been a Co-Coordinator of the Munich Data Science Research School, Oberschleißheim, Germany. Since 2019, she also heads the Helmholtz Artificial Intelligence, Weßling, with the research field "aeronautics, space, and transport." Since May 2020, she has been the Director of the International Future AI lab "AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond," Munich. Since October 2020, she has been serving as a Co-Director of the Munich Data Science Institute (MDSI), TUM. She is a Visiting AI Professor with ESA's Phi-Lab, Frascati, Italy. Her main research interests are remote sensing and Earth observation, signal processing, machine learning, and data science, with a special application focus on global urban mapping.

Dr. Zhu is a member of the Young Academy (Junge Akademie/Junges Kolleg) at the Berlin-Brandenburg Academy of Sciences and Humanities and the German National Academy of Sciences Leopoldina and the Bavarian Academy of Sciences and Humanities. She is an Associate Editor of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. She also serves as an Area Editor responsible for special issues of *IEEE Signal Processing Magazine*.

A.2 A Deep Active Contour Model for Delineating Glacier Calving Fronts

Reference

K. Heidler, L. Mou, E. Loebel, M. Scheinert, S. Lefèvre, and X. X. Zhu, "A deep active contour model for delineating glacier calving fronts," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–12, 2023. DOI: 10.1109/TGRS.2023.3296539

Copyright

A Deep Active Contour Model for Delineating Glacier Calving Fronts

Konrad Heidler, *Student Member, IEEE*, Lichao Mou, Erik Loebel, Mirko Scheinert, Sébastien Lefèvre, *Senior Member, IEEE*, and Xiao Xiang Zhu, *Fellow, IEEE*

***Abstract*— Choosing how to encode a real-world problem as a machine learning task is an important design decision in machine learning. The task of the glacier calving front modeling has often been approached as a semantic segmentation task. Recent studies have shown that combining segmentation with edge detection can improve the accuracy of calving front detectors. Building on this observation, we completely rephrase the task as a contour tracing problem and propose a model for explicit contour detection that does not incorporate any dense predictions as intermediate steps. The proposed approach, called "Charting Outlines by Recurrent Adaptation" (COBRA), combines convolutional neural networks (CNNs) for feature extraction and active contour (AC) models for delineation. By training and evaluating several large-scale datasets of Greenland's outlet glaciers, we show that this approach indeed outperforms the aforementioned methods based on segmentation and edge-detection. Finally, we demonstrate that explicit contour detection has benefits over pixel-wise methods when quantifying the models' prediction uncertainties. The project page containing the code and animated model predictions can be found at https://khdlr.github.io/COBRA/.**

***Index Terms*— Active contours (ACs), edge detection, glacier front, Greenland, uncertainty.**

I. Introduction

RECENT years have seen rapid warming in the polar regions, which has led to an exceptionally large mass loss of the Greenland ice sheet [1]. This loss of ice mass translates into global sea level rise and can cause feedback effects that further increase the warming of the Arctic [2]. Closely monitoring the Greenland ice sheet is therefore of paramount importance. About half of this ice mass loss is generally attributed to glacier dynamics, like dynamic imbalance and increased discharge. The remaining half is attributed to negative surface mass balance, which mostly stems from an increase in surface melt [3].

Following the rapid changes in air and sea temperature, glacier dynamics in these regions are changing quickly. One essential indicator for dynamic changes of marine-terminating glaciers is the calving front, which is the boundary line of the glacier from which ice bergs calve off. In order to better understand the glaciological processes and provide more accurate constraints for glacier modeling, detailed monitoring of the glaciers' calving fronts is necessary. With the ever-growing availability of satellite remote sensing data, monitoring glaciers at a large scale with high temporal frequency has become possible, but requires automated methods. Therefore, recent years have seen rapid advances in applying machine learning for glacier monitoring, which will be explored in more detail in Section II-A.

With the rise of deep learning methods in remote sensing, the predominant method of approaching this task has been via *semantic segmentation*. In this formulation, each pixel is assigned a label that corresponds to either the *glacier* class or the *sea* class. Given the large number of studies on semantic segmentation in computer vision, the methods and models for this task are well understood and provide decent results when applied to calving front detection. However, these methods require postprocessing steps to extract the actual calving front from the segmentation masks.

Noting that segmentation is only a proxy for the actual task of calving front detection, and neither the sea nor the inward glacier area is of actual interest for calving front detection, the field has seen a recent trend toward edge detection methods. By combining computer vision methods for pixel-wise edge detection with the aforementioned segmentation task, predictions are thus greatly improved [4], [5].

The goal of this study is to provide a new angle on this task. Picking up the trend toward edge detection, we propose to completely move away from pixel-wise prediction architectures and rethink the task from the ground up. The desired final prediction format for calving front detection is a vectorized polyline, which is a data format that is well-suited for downstream analysis and modeling applications. Therefore,

Manuscript received 17 November 2022; revised 11 May 2023; accepted 17 June 2023. Date of publication 27 July 2023; date of current version 2 August 2023. The work of Konrad Heidler, Lichao Mou, and Xiao Xiang Zhu was supported in part by the German Federal Ministry of Education and Research (BMBF) within the framework of the International Future AI Laboratory "AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics, and Beyond" under Grant 01DD20001; and in part by the German Federal Ministry for Economic Affairs and Climate Action within the framework of the "National Center of Excellence ML4Earth" under Grant 50EE2201C. The work of Erik Loebel and Mirko Scheinert was supported by the Helmholtz Association through the Helmholtz Information and Data Science Incubator Project "Artificial Intelligence for Cold Regions" (AI-Core). An earlier version of this paper was presented at the International Geoscience and Remote Sensing Symposium (IGARSS) 2022 [DOI: 10.1109/IGARSS46834.2022.9884819]. *(Corresponding authors: Xiao Xiang Zhu; Lichao Mou.)*

Konrad Heidler, Lichao Mou, and Xiao Xiang Zhu are with the Chair of Data Science in Earth Observation (SiPEO), Department of Aerospace and Geodesy, School of Engineering and Design, Technical University of Munich (TUM), 80333 Munich, Germany (e-mail: k.heidler@tum.de; lichao.mou@tum.de; xiaoxiang.zhu@tum.de).

Erik Loebel and Mirko Scheinert are with the Institut für Planetare Geodäsie, Technische Universität Dresden, 01069 Dresden, Germany (e-mail: erik.loebel@tu-dresden.de; mirko.scheinert@tu-dresden.de).

Sébastien Lefèvre is with IRISA UMR 6074, Université Bretagne Sud, 56000 Vannes, France (e-mail: sebastien.lefevre@univ-ubs.fr).

Digital Object Identifier 10.1109/TGRS.2023.3296539

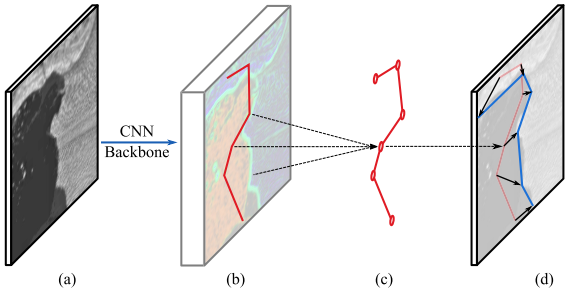


Fig. 1. High-level overview of our deep AC model for delineating calving fronts. (a) First, the backbone network takes the input image and derives feature maps. (b) Then, a sample is taken from these feature maps at the position of each vertex. (c) These features are evaluated by the Snake Head which predicts offsets for each vertex. (d) Finally, the offsets are applied to update the contour. This process is repeated multiple times.

we are looking to build a model that directly outputs the calving fronts in this desired format instead of recovering the vectorized contour from intermediate predictions. By radically redesigning the neural network architecture, we are able to move away from pixel-wise classifiers and instead arrive at a model that directly predicts the calving front as a polyline.

This approach has several theoretical benefits over representing the desired output by a dense, pixel-wise mask.

1) As the predictions are already in a vector format, there is no need for complicated postprocessing pipelines like with pixel-wise approaches.
2) By its very design, the model will learn to focus on the actual object of interest, the calving front.
3) Looking closer into the application, explicit contour prediction provides a natural way of encoding prior knowledge into the network. In pixel-wise detection frameworks, the network may predict undesired outputs like disconnected line segments. By directly predicting an explicit contour, such issues are eliminated.
4) Explicit contours make more efficient use of computational resources. A sequence of vertices takes fewer parameters than a dense mask.
5) Finally, the vectorized representation allows for better quantification of model uncertainty as the joint probability distribution of a sequence of vertices is easier to model than that of pixel-wise masks.

Convinced by these theoretical considerations, we set out to develop a calving front detection model that directly predicts the desired contours, as shown in Fig. 1. Contour-based approaches for the segmentation of regions in natural images have been extensively studied in the form of *active contours*, which are also called *Snakes* [6].

In order to provide robust and stable calving front predictions for downstream applications, such as glaciological studies and models, it is important to quantify the reliability of the model's predictions. Therefore, we also explore the question of uncertainty quantification in calving front detection. Experimental results using the Monte Carlo (MC) dropout method [7] across different models suggest that uncertainty quantification with contour-based models can indeed bring benefits over pixel-wise models.

Overall, we summarize the goals and contributions of this study by the following points.

1) We rephrase the task of the automated glacier calving front detection from a segmentation task to a contour detection task and show that deep active contour (AC) models are a feasible approach to solving this task.
2) We develop a specialized deep AC model for the delineation of glacier calving fronts which outperforms both pixel-based approaches as well as existing deep AC models. The effect of the design decisions is validated through extensive ablation studies.
3) We explore the benefits of contour-based methods for uncertainty quantification compared to pixel-wise methods.

II. Related Work

In order to place our work into the context of existing research, we provide a brief overview of existing calving front detectors, as well as methods for explicit edge predictions.

A. Detecting Calving Fronts in the Deep Learning Era

Given the strong performance of deep learning-based methods for calving front detection, traditional vision methods have largely become insignificant for this task [5], [8]. Therefore, this section focuses on deep learning-based methods. Here, most approaches formulate the task as a variant of sea–land segmentation. In this formulation, a semantic segmentation network is used to separate the image into land and ocean classes [8]. There is a considerable number of well-tested segmentation architectures, like UNet [9], which is a strong baseline for most segmentation tasks. Even without any changes to the network itself, this approach can yield satisfactory results for calving front detection, which has been shown in previous studies for both the Greenland ice sheet [10] and the Antarctic ice sheet [11]. Seeing this strong baseline performance of the UNet, Periyasamy et al. [12] show that the performance of such a model can be greatly improved by tweaking network components, like normalization layers, the loss function, or dropout rate.

Further progress in this field was made by extending the UNet model or exploiting the advances of more recent segmentation model architectures. For example, Loebel et al. [13] add more layers and thereby increase the number of down- and upsampling steps. This enlarges the spatial context that the network can consider for its decisions and therefore leads to better predictions. Following recent advances in transformer-based model architectures, Holzmann et al. [14] enhance the UNet model with attention gates to improve the interpretability of the model and better understand the learning process. Another newer neural network architecture that has successfully been adopted for calving front detection is DeepLabv3+ [15]. Zhang et al. [16], Cheng et al. [4], and Gourmelon et al. [17] bring in ideas from this architecture to obtain more accurate delineations of glacier calving fronts. Finally, it is also possible to combine image classification and segmentation [18], which can lead to more robust results.

Recently, there appears to be a trend toward models that approach calving front detection by extending or even replacing semantic segmentation with edge detection methods.

By focusing on the boundary between the two classes rather than the areas of sea and glacier, these models are encouraged to learn features that are informative of the calving front rather than features of the sea and glacier areas.

Wavelet transforms are one possible approach that makes use of the abrupt changes in texture between glacier and sea to determine the location of the calving front [19]. Davari et al. [20] use a different transformation, namely the Euclidean distance transformation, and train a network that predicts each pixel's distance to the calving front instead of a binary class. Great potential lies especially in the combination of segmentation and edge detection. Both HED-UNet [5] and the calving front machine (CALFIN) [4] choose this approach to outperform models that focus on only one of these two aspects.

Contrary to these models, our approach is not to predict pixel-wise segmentation or edge masks, but instead to explicitly predict a contour parameterized by a fixed number of vertices.

B. Explicit Edge Prediction

The idea of explicitly parameterizing contours in an image was pioneered quite early in the history of computer vision by Kass et al. [6]. They proposed *active contours* or *Snakes*, which evolve from an initial contour by iteratively minimizing an energy functional. By design, this functional takes its minimum when the contour coincides with the desired boundary in the image. Using Radarsat data, this approach has been shown to work for the delineation of the Antarctic coastline on a coarse scale [21]. The main drawback of conventional AC methods is the fact that they are limited to single-channel imagery without any natural extension to multichannel imagery. Furthermore, they are sensitive to local image contrast and the results depend highly on the initialization of the contour.

As automatic feature extraction is a strong suit of deep learning models, the idea of combining ACs with deep learning is not a new one. Rupprecht et al. [22] introduced a deep AC model that works by first predicting a 2-D offset field that points from each pixel toward the closest boundary point. An initial contour will then evolve along this offset field until it converges. However, this method is not end-to-end trainable as it relies on the intermediate offset field and no gradients flow through the actual curve evolution. While this approach can work on multichannel imagery, it still suffers from a strong dependence on contour initialization.

In an effort to introduce an end-to-end trainable deep AC model, Peng et al. [23] proposed to make not only the feature extraction part learnable, but the contour evolution step as well. Their model, termed *Deep Snake* first derives feature maps using a convolutional backbone network and then samples the features at the position of each vertex. From these sampled features, a 1-D convolutional neural network (CNN) then predicts the offsets for each vertex. Like with conventional AC models, this process is then iterated to refine the predictions.

As one of the most recent models in this line of research, deep attentive contours (*DANCE*) [24] improves on the Deep Snake idea by introducing an "edge attention map," which influences the speed of the snake evolution. While vertices far from the target boundary are evolving quickly, the update speed for points closer to the boundary is slowed down.

Inspired by these advances, our goal is to develop an AC model for the task of calving front detection. The existing models address the computer vision task of instance segmentation, where objects in an image are locally segmented. For calving front delineation, however, one global line between glaciers and the ocean is needed. This disparity and further differences, like the general shapes of the objects of interest, call for a completely different network architecture as well as changes to loss functions and the network training protocol.

III. Deep Active Contour Models for Calving Front Delineation

When approaching the task of calving front detection, we first take a look at how a human would proceed in solving the task. In discussions with experts and when annotating calving fronts ourselves, one central observation is the order in which different areas in the scene are addressed.

To a human annotator, it does not make much difference whether they are told to trace the edge between two objects in an image or fill in the areas that both objects occupy. In both cases, they will usually start out by tracing the boundary area between the two classes with minute attention to detail. When asked to perform segmentation, the remaining areas are then filled in with broad strokes in a second step. So while humans will approach both tasks in a similar fashion, a focus on the boundary appears to be the more natural way to formulate the task of calving front detection.

For a neural network model, the way a task is formulated changes everything. As these models are trained to minimize some loss function, the final performance is determined by how well a low loss value translates to good performance on the actual task. For instance, it has been observed that the cross-entropy loss used for training semantic segmentation models can lead to blurry edges between the classes due to the fact that each pixel contributes equally to the final loss, no matter its position in relation to the objects in the image. This implies that a model can minimize most of its loss by correctly classifying the simpler pixels that lie in the interior of the objects of interest. In turn, the model will spend less attention on the pixels near the boundary, which are much more important for solving the task [5], [25].

Phenomena like these are likely the reason that calving front delineation has recently seen a shift toward improving these segmentation models by including pixel-wise edge detection tasks. By putting more focus on the edges, the model is forced to learn how to better distinguish the classes in these critical areas [4], [5]. Instead of combining segmentation with edge detection in a pixel-wise framework, we take a more radical approach in this study. By completely eliminating the semantic segmentation aspect and focusing only on the edges, it is possible to reformulate the task in such a way that it does not require pixel-wise classifications. Instead, the model will directly output a vectorized contour.

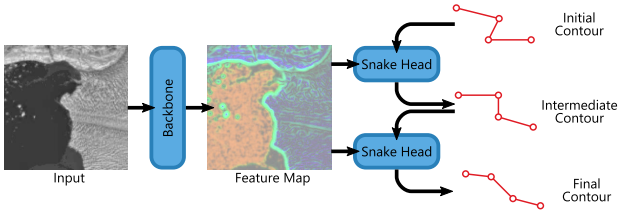


Fig. 2. Architecture overview of our model. Note that while this diagram shows only two iterations of the Snake Head, the number of iterations is actually an arbitrary hyperparameter, which we set to four for our experiments.

A. General Model Architecture

Inspired by the ideas of Peng et al. [23], we develop a deep contour model for the delineation of glacier calving fronts. As is common in many recent computer vision models, our model consists of two main components which perform different subtasks in order to solve the overall task together. The *backbone* is a general-purpose 2-D CNN which is used to extract semantically valuable features from the input imagery. The second component is a *prediction head*, which makes use of the backbone's features to derive the final network predictions. In our model, the prediction head takes the role of the AC iteration. Therefore, it will take a contour and the backbone's feature maps as its inputs, and update the contour to better match the desired boundary. Due to this functional similarity, we call this component the *Snake Head*. The overall architecture of the network is visualized in Fig. 2. Notably, this framework can be trained end-to-end, as all components are fully differentiable.

For the backbone, multiple feature extractors were evaluated. Initial experiments with standard ResNet backbones produced unsatisfactory results. This leads us to believe that while ResNets are a strong backbone for many vision tasks, they are likely not optimal for deep AC models. In search of a better-suited backbone network, the Xception backbone [26] used by Cheng et al. [4] in their study of Greenland's glaciers proved to be a very capable backbone for remote sensing of glaciers that transfers well to deep AC models.

B. Snake Head

The central challenge in predicting contours from an image is the fact that input and output are represented in different dimensionalities. While the input image is represented by a 2-D grid of pixels, the contour that the model should output is given as a sequence of vertices, which is 1-D. The idea of AC models is to start with an initial contour and then iteratively update this contour based on the image values at each vertex. Conceptually, deep AC models do nearly the same thing. However, they do not directly sample the image values but instead, sample the values from the feature map derived by the backbone network.

After sampling the backbone features at the vertex positions, the Snake Head predicts an offset for each vertex, which represents how the vertex needs to be shifted so that the entire sequence of vertices can better represent the true contour. This is achieved by using a 1-D CNN. While conventional, 2-D CNNs pass information between adjacent pixels, the 1-D CNN used in the Snake Head passes information between adjacent vertices of the contour. In order to enable the passing of information between vertices that are far away from each other, we stack multiple such convolutional layers. The receptive field of the Snake Head is further increased by using dilated convolutions. In our model, the Snake Head is therefore given as a stack of dilated convolutions. We set the sequence of dilation rates to 1, 3, 9, 9, 3, and 1, as similar setups have proven to be successful at capturing low- and high-frequency features in signal processing tasks [27].

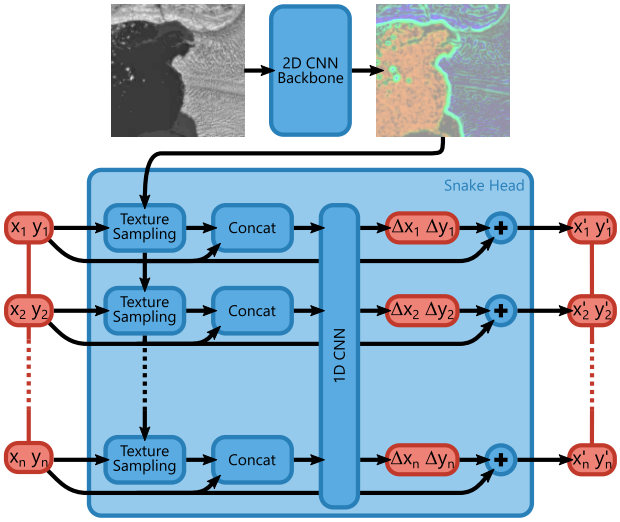


Fig. 3. Detailed view of the Snake Head. After the feature maps are sampled at the vertex positions, the vertex coordinates are concatenated to the vertex features. The 1-D CNN then predicts offsets for each vertex. These offsets are added to the input coordinates to obtain the Snake Head's output.

In order for the Snake Head to gain some spatial reasoning capabilities, we also include the vertex coordinates as additional input features for the Snake Head CNN. This allows the model to not only ensure a homogeneous spacing of the output vertices but also to learn some prior assumptions on the shape of the calving fronts. The overall working mechanism of the Snake Head is shown in Fig. 3.

To translate the iterative nature of the AC method, we apply the Snake Head multiple times with the shared weights to obtain more refined predictions. Starting with the initial contour, the Snake Head samples features at the vertex positions, calculates and applies the offsets, and repeats the process. Compared to conventional AC models, which can take dozens [21] or even hundreds [28] of iterations to converge, the deep AC model converges to satisfactory results after a small number of iterations. For our experiments, we set the number of iterations to four.

In the context of deep neural networks, the Snake Head can also be regarded as a recurrent neural network. The locations of the vertices then represent the hidden state of the network, which is updated throughout the iteration steps until the final output is achieved.

C. Loss Function

The loss function is a crucial element of any deep learning model, as it measures how well the model is performing

and gives feedback for improving the network via backpropagation. When predicting contours, the loss function should therefore measure the similarity between the predicted contour p, represented by vertices p_i with $1 \leq i \leq V$, and the true contour t, given by the vertices t_j with $1 \leq j \leq V$.

Common loss functions for polygon regression are based on the L_1 and L_2 losses, which, following the above notation, are defined as follows:

$$L_1(p,t) = \frac{1}{V}\sum_i \|p_i - t_i\| \tag{1}$$

$$L_2(p,t) = \frac{1}{V}\sum_i \|p_i - t_i\|^2. \tag{2}$$

These loss functions have a fundamental issue when predicting glacier front lines. By only computing the distance between the vertices of p and t at the same index, they tacitly assume that each predicted vertex corresponds to exactly one vertex in the true contour. However, the model has no way of knowing how the ground truth vertices were placed along the true contour. In the setting of Deep Snake and DANCE, the vertices were placed equidistantly along the contour of objects that were largely convex, so this assumption did not have much negative impact there.

When predicting glacier frontlines, however, this issue becomes much more prominent due to the irregular and jagged shape of these contours. As the model essentially tries to minimize the L_1 or L_2 loss for a number of possible parameterizations of the true contour at the same time, the resulting predictions lack sharp edges and instead follow a smoothed version of the true outline.

Naturally, contour prediction is not the first task to face challenges like these. For example, in the context of time-series analysis, slight variations in timing are often less important than the general shape of the time-series. *Dynamic time warping* (DTW) is a method that was proposed by Sakoe and Chiba [29] in order to address this very issue. Given two sequences, they not only compare the pairwise differences but instead, first, find an optimal alignment between the two sequences and then calculate the distances based on that alignment.

In our setting, the parameterization of a contour takes the role of time in the original DTW. Formally, we define the DTW loss for two contours p and t to be

$$\mathcal{L}_{\text{DTW}}(p,t) = \min_{(i_k,j_k)_{k\in[K]}\in\mathcal{K}} \sum_k \|p_{i_k} - t_{j_k}\|_2^2 \tag{3}$$

where $\mathcal{K}$ denotes the set of all possible realignments $(i_k, j_k)_{k\in[K]}$ that satisfy the following three conditions.

1) For any $i \in \{1,\ldots,V\}$, there is a k with $i_k = i$.
2) For any $j \in \{1,\ldots,V\}$, there is a k with $j_k = j$.
3) The sequences i_k and j_k are nondecreasing in k.

Under these conditions, the DTW loss can be efficiently calculated using dynamic programming [29].

A possible issue with the use of DTW as a loss function in deep learning is the fact that it is not smooth due to the minimum operator applied in (3). Seeing this, Cuturi and Blondel [30] replace the minimum with a *soft minimum* which they define as

$$\text{softmin}_\gamma(x_1,\ldots,x_n) = -\gamma \log \sum_{k=1}^{n} \exp\frac{-x_k}{\gamma} \tag{4}$$

with a smoothness parameter $\gamma > 0$. In the limit $\gamma \to 0$, the conventional minimum operator is recovered.

D. Implementation Details

A central issue with naively backpropagating the loss through the snake iteration is the fact that the early iterations show poor convergence to the target contour. This is easily fixed by stopping the gradient from flowing through the coordinates at the beginning of each snake step. To still encourage quick convergence of the contours during inference, we leverage deep supervision by including an additional loss term for each intermediate step. During training, the current contour is compared to the ground truth after each snake step, and the resulting loss is added toward the final loss for the gradient calculation. Unless otherwise stated, all models use a contour parametrization by 64 vertices.

All models in the study were trained for 500 epochs on the training dataset. We used the Adam optimizer [31] with an initial learning rate of 10^{-3} decaying to $4 \cdot 10^{-5}$ on a cosine decay schedule [32].

Our models are implemented in JAX [33] using the Haiku framework [34]. The training was conducted on a single NVIDIA RTX 3090 GPU with 24 GB of VRAM. Training an instance of the model, took around 25 h, and had an estimated energy consumption of 8.1 kWh.

IV. Experiments and Results

A. Datasets

In order to thoroughly evaluate our model and compare it with other approaches, we choose two large-scale datasets of marine-terminating glaciers in Greenland for training and evaluation purposes, namely the *CALFIN* dataset [4] and the calving front dataset from *TU Dresden* (TUD) [13]. Both of these datasets include respective testing data. Furthermore, the Baumhoer dataset [11] consists of synthetic aperture radar (SAR) data of Antarctic glaciers, thus serving as a benchmark for the models' ability to generalize to a different data modality and a different ice sheet.

1) CALFIN Dataset: The CALFIN dataset consists of near-infrared data from the various Landsat missions and is most notable for the long time span of acquisition times, ranging from 1972 to 2019. Its spatial coverage is 66 Greenlandic glaciers, which amount to 1541 Landsat scenes. In an effort to improve generalizability to different sensors, the training dataset also includes 232 single-polarization Sentinel-1 scenes of glaciers in Antarctica. The corresponding test dataset consists of 162 Landsat near-infrared scenes. For all of the mentioned scenes, the calving fronts were manually delineated.

2) TUD Dataset: In contrast to this, the TUD dataset puts its focus on the eighth iteration of the Landsat mission, providing a dense time-series of recent acquisitions of Greenland's marine-terminating glaciers. The captured scenes

range from 2013 to 2021 for a total of 1127 tiles. For studies related to feature importance and data fusion, it includes the full multispectral imagery, as well as topography data and texture information derived using gray-level co-occurrence matrix statistics. For interoperability with the other datasets, only the panchromatic imagery is used in this study.

3) Baumhoer Dataset: Another test set for evaluating the generalization of the trained models is given by the Baumhoer dataset [11]. This dataset is vastly different from the other datasets, as the imagery is not from Greenland, but from Antarctica instead. Furthermore, it consists of Sentinel-1 SAR imagery, which marks a second challenge in generalization. While the original dataset is not openly available, an evaluation subset is distributed along with the CALFIN dataset [4]. In order to keep this study fully reproducible, we only use this publicly available subset of the Baumhoer dataset. The used testing set consists of 62 Sentinel-1 scenes of glaciers in Antarctica from the year 2018.

B. Evaluation Metric

As there is no uniquely defined distance metric between two curves, many different metrics are being used for evaluating the accuracy of predicted glacier frontline positions [8]. In our work, we adapt the Polis metric [35], which was originally proposed for measuring the dissimilarity between building footprints. For two polylines v and w, with I and J vertices, respectively, it is defined as the average distance of any vertex to the respective other polyline

$$p(v, w) = \frac{1}{I}\sum_{i=1}^{I} d(v_i, w) + \frac{1}{J}\sum_{j=1}^{J} d(w_j, v)$$

where $d(v_i, w)$ denotes the distance between vertex v_i and the closest point on the polyline w. Note that this closest point w does not need to be a vertex, but may be a point between vertices as well.

Compared to other existing metrics, like the Fréchet distance [36], which is defined as the solution to a min-max problem, the Polis metric is more easily interpretable as the "average" distance between the two curves. Furthermore, it was chosen due to its symmetry and the fact that it takes all predicted points into consideration.

C. Comparison With Other Models

For our comparison study, we train a number of different models in order to compare their performance on the test datasets. To compare with the state of the art in calving front detection and contour-based outline detection, we include both pixel-wise and contour-based models.

1) Pixel-Wise Models: This first group of methods consists of pixel-wise segmentation models that are known to work well for calving front detection.

a) UNet [9]: This model is a popular semantic segmentation model that serves as a strong baseline for many segmentation tasks. It has been successfully applied to calving front detection [11].

b) DeepUNet [13]: The model developed and used by Loebel et al. [13]. Its main difference from the original UNet model is the addition of two down- and upsampling steps, which make the model deeper and more aware of spatial context.

c) HED-UNet [5]: A combination of the UNet model with an edge detection model, HED-UNet was originally developed to detect glacier frontlines on the Baumhoer dataset [11].

d) CALFIN [4]: This model was introduced by Cheng et al. [4]. The model is based on the segmentation architecture DeepLabv3+ [15]. It is the first to leverage the potential of the Xception network for calving front detection.

2) Contour-Based Models: For a comparison to existing contour-based models, we also include models from this group in the comparison. It should be noted that unlike the pixel-wise models above, they were not developed for calving front detection.

a) Deep AC [22]: One of the first works to combine AC models with deep learning, this model uses a 2-D CNN to predict an offset field that points toward the nearest contour point from each pixel and then evolves a contour along this offset field.

b) Deep Snake [23]: Originally proposed as a contour-based model for instance segmentation, we have made slight changes to this model to perform calving front detection. Specifically, the circular convolutions in the network were replaced with regular 1-D convolutions, as the predicted contours for calving fronts should be open polylines and not closed polygons. Furthermore, the object detection head of the model was removed as with calving fronts, there is always exactly one contour to be predicted. For a fair comparison, we train and evaluate this model not only with the ResNet-50 backbone but also with the Xception backbone.

c) DANCE [24]: An iteration of the Deep Snake [23] model, DANCE introduces an edge attention map that speeds up the evolution for vertices far from the true edge and slows the evolution for vertices on the true edge. We applied the same adaptations to this model as to the Deep Snake model.

As CALFIN provides the largest and longest record of glacier observation, we train all models on the CALFIN training set and then evaluate them on the CALFIN, TUD, and Baumhoer test sets. The numerical evaluation results for this comparison study are displayed in Table I, and some visual results are displayed in Fig. 4. For the proposed Charting Outlines by Recurrent Adaptation (COBRA) model, we train three randomly initialized models and report mean and standard deviation across these three runs.

Comparing the pixel-wise models, we can reproduce the increased performance of the calving front-specific models over the baseline UNet. All three of these models, namely DeepUNet, HED-UNet, and the CALFIN cut down the average prediction error from UNet's 224 m to the range of 130–138 m on the CALFIN test set. There is, however, a difference in generalization to the other datasets, where the DeepUNet seems to generalize best to SAR imagery, and the CALFIN generalizes better than others on the TUD dataset, which is also based on Landsat imagery.

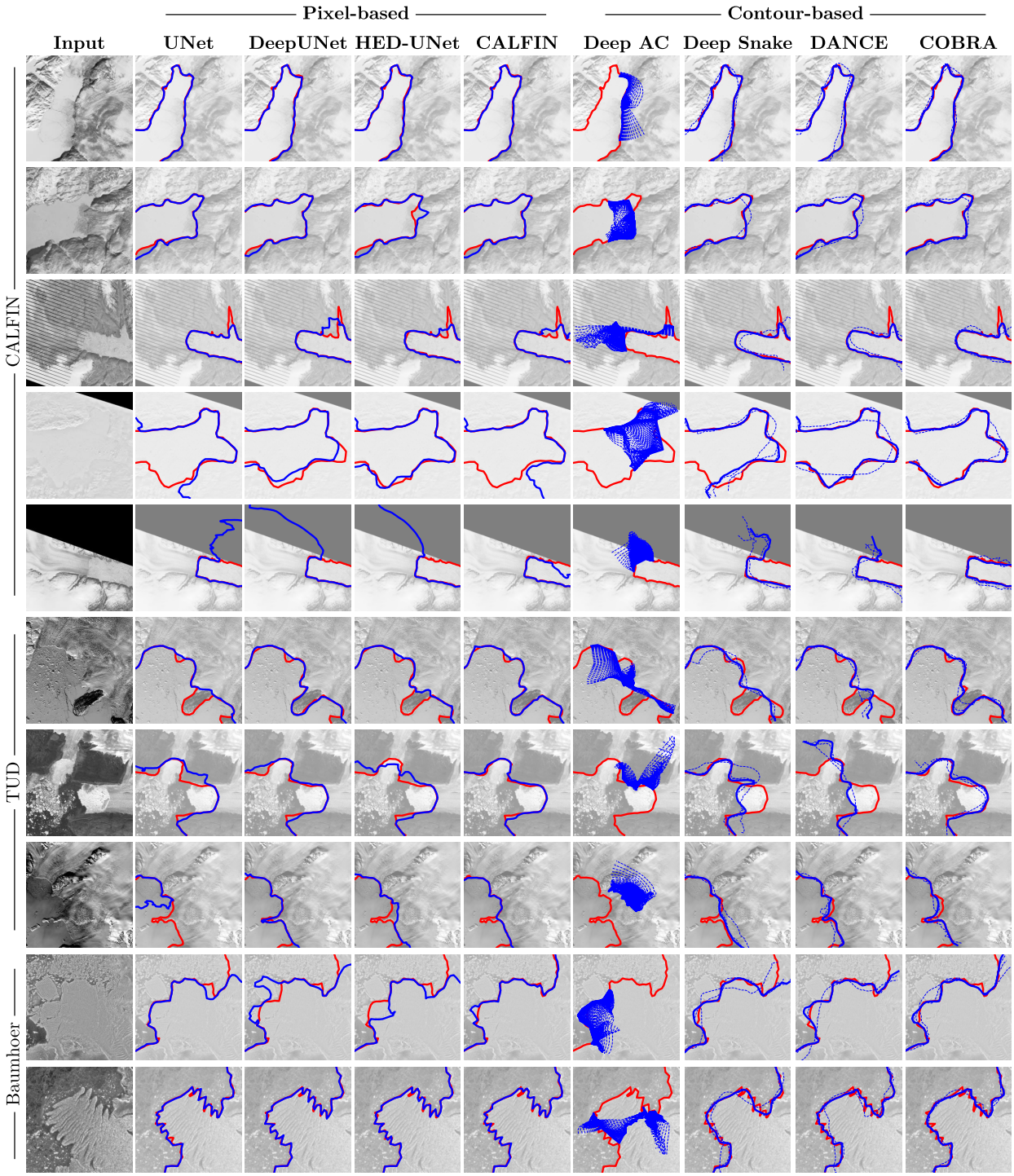


Fig. 4. Visualization of prediction results (blue) for the different models and corresponding ground truth (red) on the test datasets. For the iterative, contour-based models intermediate results are displayed as blue dashed lines. Best viewed in color.

Looking at the contour-based models, the Deep AC model falls behind the competition and performs the worst out of all the models in our experiments. The bad performance of the Deep AC model is easily explained when looking at the visual results in Fig. 4. It stems from the fact that its predictions tend to only represent a part of the desired curve, as there is no regularization term that forces the prediction to cover the entire calving front.

With the exception of the Deep AC model [22], the contour-based approaches perform considerably better on the CALFIN evaluation than their pixel-based counterparts, especially when using the Xception backbone network.

On the other hand, generalizing to the Antarctic Baumhoer dataset is particularly hard for contour-based methods. We attribute this to the presence of jagged floating ice-tongues (see Fig. 4, last row), which are not observed in the same

TABLE I
MEAN DEVIATION (POLIS METRIC) OF THE TRAINED MODELS ON THE EVALUATED TEST SETS

	CALFIN	TUD	Baumhoer
UNet [9]	224 m	288 m	122 m
DeepUNet [13]	138 m	241 m	92 m
HED-UNet [5]	130 m	231 m	122 m
CALFIN [4]	138 m	159 m	99 m
Deep Active Contours [22]	515 m	699 m	652 m
Deep Snake RN50 [23]	289 m	467 m	247 m
Deep Snake Xception [23]	123 m	316 m	130 m
DANCE RN50 [24]	118 m	290 m	131 m
DANCE Xception [24]	103 m	272 m	102 m
COBRA (mean of 3)	99 m	144 m	99 m
COBRA (standard deviation)	±10 m	± 21m	±12m

TABLE II
UNCERTAINTY QUANTIFICATION: PEARSON CORRELATION BETWEEN MODEL UNCERTAINTY AND ACTUAL PREDICTION ERROR (POLIS METRIC)

Dataset	CALFIN	TUD	Baumhoer
UNet [9]	0.3603	0.3864	0.1873
DeepUNet [13]	0.2762	0.3726	0.4191
HED-UNet [5]	0.3391	0.5518	0.3989
CALFIN [4]	0.1790	0.3449	0.2337
Deep Active Contours [22]	0.0842	0.2190	0.1759
Deep Snake Xception [23]	0.2009	0.4959	0.4508
DANCE Xception [24]	0.2972	0.3346	0.3589
COBRA	0.4811	0.4414	0.6031

way during training. Such features lead to more complex outlines, which need more vertices for their representation, as can be seen from the results of the study on vertex numbers in Section IV-E2.

Overall, our proposed network outperforms both the pixel-based and other contour-based models by a considerable margin on the CALFIN and TUD evaluations. Even when generalizing to the radically different Baumhoer dataset, our model still maintains a respectable performance.

Inference results from the COBRA model for the three testing datasets are available for online viewing and as a shapefile download at https://github.com/khdlr/COBRA/tree/master/inference_results.

D. Quantifying Uncertainty With Contour Models

With deep learning models growing ever more complex, quantifying the uncertainty of their predictions at inference time has become an important consideration when working with such models. Deep learning models being over-confident in their predictions is a common issue [37]. As the models are usually trained on definite ground truth, the models are never taught to concede their uncertainty in ambiguous cases.

One elegant method for the quantification of network uncertainties is known as MC dropout (MCD). In their seminal study, Gal and Ghahramani [7] demonstrated that a deep learning model trained with dropout layers can be interpreted as approximated Bayesian inference in a deep Gaussian process. Samples from the posterior distribution approximated by such a model can be recovered quite easily by enabling the dropout layers not only at training time but also at inference time. It has been shown that MCD can quantify model uncertainties well for remote sensing tasks like aerial image segmentation [38]. Recently, Hartmann et al. [39] also successfully applied a Bayesian UNet for the segmentation of glaciers in SAR imagery.

As the MCD method is simple to implement and evaluate compared to other uncertainty quantification methods, we choose this approach for quantifying uncertainties in the model predictions. In order to estimate the hardness of samples at inference time and get an estimate for the model uncertainty, we calculate the original, deterministic model prediction, as well as multiple additional predictions using the MCD technique. If these predictions all line up well with the original model prediction, the model can be assumed to be quite certain of its prediction. On the other hand, a large deviation between the original model prediction and the MCD predictions corresponds to ambiguity in the model output, implying a potentially higher prediction error.

Taking the ten MC samples and the model's deterministic prediction, we estimate the model uncertainty as the average Polis-distance of each MC sample from the deterministic prediction.

Our hypothesis is that the explicit edge parameterization by vertices lends itself much better to uncertainty quantification from these posterior samples than dense pixel-wise predictions, due to the fact that the explicit representation requires much fewer parameters. With fewer parameters, the covariance between the parameters becomes more tractable, and therefore easier to approximate for any model.

For our uncertainty quantification study, we apply MCD with a dropout rate of 20% to the aforementioned models. After training the models, we draw ten predictions with enabled dropout (posterior samples) per model for each test scene in order to assess the quality of the uncertainty quantification.

Fig. 5 shows the posterior samples obtained using the MCD models. It can be observed that the pixel-wise calving front detectors can collapse completely on hard scenes. All the evaluated pixel-wise methods suffer from this phenomenon, suggesting that it is indeed related to the mode of representation. Inspection of the underlying segmentation masks suggests that this is due to the fact that when working with segmentation masks, small changes in the segmentation can completely change the topology of the prediction as previously connected regions can become disconnected and vice-versa. Due to this effect, estimating the model uncertainty using MCD can overestimate the hardness of the samples for these models on easy scenes.

By their design, contour-based methods do not have this limitation, as they predict the frontline directly. Among these models, the DANCE architecture seems affected by similar issues as the pixel-wise models, which we attribute to the fact that DANCE incorporates an intermediate dense prediction. Overall, both Deep Snake [23] and our proposed model appear to be best suited for uncertainty quantification using MCD, with very similar samples on easy scenes, and deviating

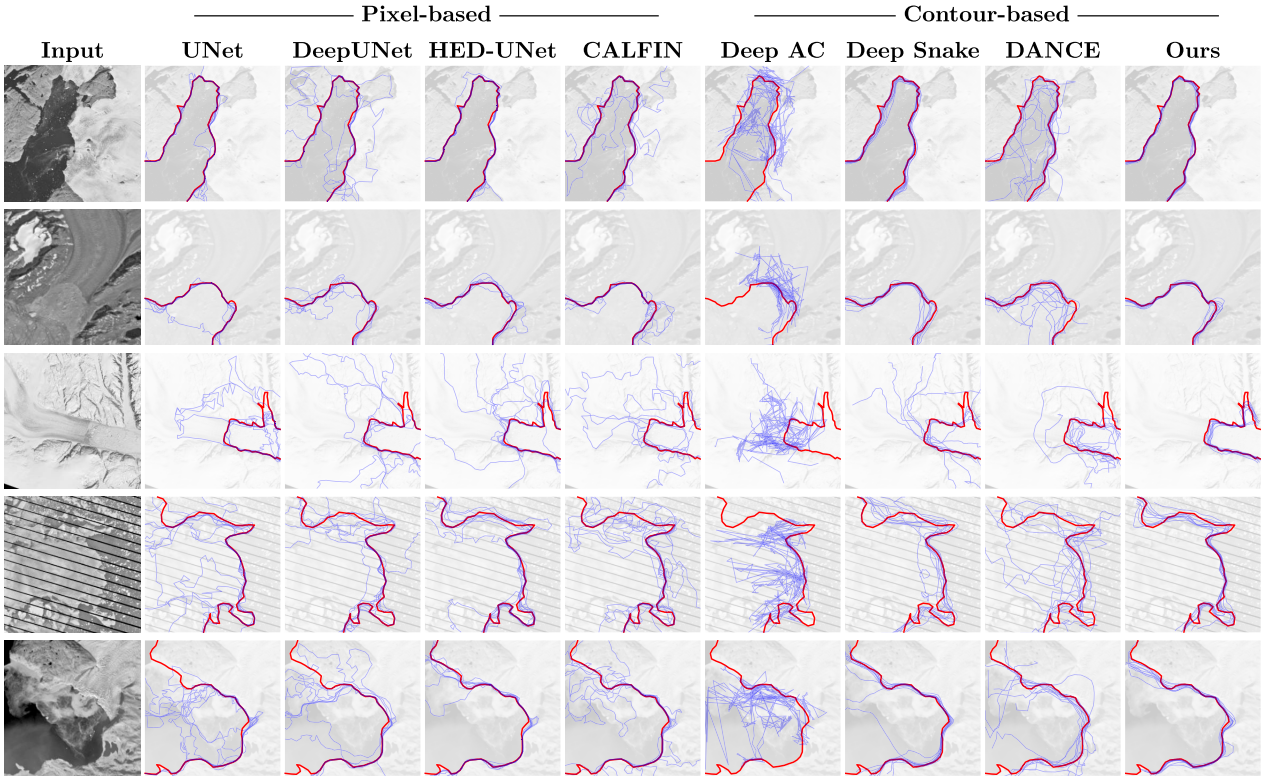


Fig. 5. Visualization of posterior samples obtained using MCD (blue) from the different models overlaid on top of the ground truth (red) for scenes from the CALFIN test set.

samples in areas that are harder to delineate. Among the pixel-based methods, HED-UNet [5] appears to be the best at quantifying its uncertainty.

In an effort to numerically evaluate the uncertainty quantification, we then calculate the Pearson correlation coefficient between the model uncertainty and the actual prediction error. A high correlation between these two variables corresponds to better uncertainty quantification, as the model should only be certain when its prediction is actually correct, while a high model uncertainty should be indicative of the prediction possibly being far from the ground truth.

The results of this evaluation are displayed in Table II. Our model is leading the evaluation for the CALFIN and Baumhoer datasets, reaching respectable Pearson coefficients of 0.4811 and 0.6031, respectively. On the TUD dataset, HED-UNet and the classic Deep Snake outperform our model on the uncertainty benchmark, achieving Pearson coefficients of 0.5518 and 0.4959. Still, our model scores decently with a Pearson coefficient of 0.4414.

These observations support our hypothesis that the contour representation is better suited for uncertainty quantification.

E. Ablation Studies

In order to better understand how the design decisions help our proposed model to improve upon the existing contour-based methods, we conduct a number of ablation studies to quantify the value of the network's components.

1) Loss Function: The rationale for implementing SoftDTW loss for our model was the assumption that the model cannot always correctly guess the placement of the vertices along the ground truth contour. Indeed, we observe better performance when using a time-warping loss, as can be seen in Table III. Surprisingly, the difference between DTW and its smooth SoftDTW variant is rather small, which suggests that the theoretical advantage of SoftDTW's smoothness does not matter much in practice for this application.

TABLE III
RESULTS OF THE LOSS FUNCTIONS ABLATION STUDY. DEVIATIONS CALCULATED USING THE POLIS METRIC

Loss Function	CALFIN	TUD	Baumhoer
L_2	114 m	312 m	119 m
L_1	102 m	296 m	111 m
DTW	90 m	236 m	91 m
SoftDTW	99 m	144 m	99 m

TABLE IV
RESULTS OF THE STUDY ON THE NUMBER OF VERTICES. DEVIATIONS CALCULATED USING THE POLIS METRIC

Vertices	16	32	64	128	256
CALFIN	120 m	98 m	99 m	98 m	116 m
TUD	252 m	209 m	144 m	214 m	261 m
Baumhoer	128 m	102 m	99 m	85 m	128 m

2) Number of Vertices: When choosing the number of vertices to represent the contours, a balance needs to be taken between too few vertices and too many vertices. Too low a number of vertices will not allow the model to sufficiently approximate the true contour, while too many vertices should lead to overfitting and issues in communicating information

TABLE V
RESULTS OF THE STUDY ON THE NUMBER OF ITERATIONS. DEVIATIONS CALCULATED USING THE POLIS METRIC

Iterations	2	3	4	5	6	7
CALFIN	193 m	147 m	99 m	157 m	152 m	157 m
TUD	318 m	262 m	144 m	271 m	267 m	278 m
Baumhoer	148 m	119 m	99 m	110 m	111 m	106 m

between vertices far apart in the sequence. In order to experimentally find a good setting for the number of vertices, we train COBRA configurations with different numbers of vertices. For computational efficiency, we always set the number of vertices as a power of two, choosing 16, 32, 64, 128, and 256 as possible vertex counts. The results of these experiments are displayed in Table IV.

For all three datasets, we observe that with increasing vertex count, performance decreases toward both ends of the tested range, which suggests that there is indeed a sweet spot around the middle of the evaluated range. For the CALFIN dataset, there appears to be an optimal performance plateau from 32 to 128 vertices while for the TUD dataset, 64 vertices are optimal. Interestingly the Baumhoer dataset seems to require a higher number of vertices for the best performance, reaching the best performance at 128 vertices. We attribute this to the aforementioned higher complexity of calving fronts in Antarctica. In practice, we recommend choosing the number of vertices accordingly to the complexity of the calving fronts of the region of interest. In general, setting it to 64 offers overall good performance across all study areas concerned in this study, the selection of which could be quite representative for large-scale applications.

3) Number of Iterations: A fundamental hyperparameter of our network is the number of iterations of the Snake Head. When given too few iterations, the model will likely not have enough capacity to converge to the right contour. On the other hand, given a large number of iterations, we expect the model to overfit the training set and generalize worse to unseen scenes. In order to find evidence for these hypotheses, we conduct a study on the number of iterations where we retrain COBRA models with iteration numbers from two to seven. The results in Table V suggest that there is indeed a sweet spot at four iterations. Starting from two iterations, performance improves considerably on all evaluation datasets up until four iterations. After that, increasing the number of iterations decreases the performance again. Therefore, we set the number of iterations for our model to four.

4) Coordinate Features: Including the vertex coordinates as additional features allows the Snake Head to take the distance and relative position of the vertices into account, but could also introduce a source of overfitting. In the ablation study (see Table VI), we observe that these coordinate features improve performance slightly on the CALFIN test set and drastically improve performance on the TUD dataset, where the average deviation is more than halved. For the Baumhoer dataset, performance degrades slightly when including coordinate features. This suggests that the coordinate features help the model to learn implicit shape priors for Greenlandic glacier calving fronts, which are not helpful when transferring the model to the Antarctic calving fronts in the Baumhoer dataset.

TABLE VI
RESULTS OF THE BINARY ABLATION STUDY. DEVIATIONS CALCULATED USING THE POLIS METRIC

Ablation	CALFIN	TUD	Baumhoer
Full Model	99 m	144 m	99 m
No Coordinate Features	95 m	301 m	100 m
No Gradient Stopping	397 m	285 m	531 m
No Deep Supervision	102 m	207 m	118 m
No Shared Weights	88 m	294 m	141 m

5) Gradient Stopping: Originally, the idea of stopping the gradients from flowing through the vertex coordinates between iterations was introduced to improve the convergence of the model. However, the "No Gradient Stopping" ablation in Table VI shows that this choice is essential for the performance of the model. Without gradient stopping, the model predictions deteriorate to a degree where they are worse than the predictions of the baseline UNet model. We attribute this to numerical instabilities in the texture sampling procedure that is used to translate between the feature maps and vertex features, which can arise from letting gradients flow through the vertex positions.

6) Deep Supervision: During training, we calculate a loss term after each snake iteration and sum up these individual loss terms for the final loss. To quantify the contribution of this deep supervision, we also evaluate a model trained without intermediate loss terms, displayed as "No Deep Supervision" in Table VI. While the in-distribution samples from the CALFIN test set do not improve much with deep supervision, generalization on TUD and Baumhoer is improved by this change.

7) Weight Sharing: The underlying hypothesis for shared weights in the Snake Head iterations was the assumption that a single set of weights would lead to better generalization results than applying a series of multiple distinct Snake Heads. The ablation results for "No Shared Weights" in Table VI support this hypothesis. On the CALFIN test set, the prediction accuracy is nearly constant between the model with shared weights and the one with distinct weights. However, on the other test sets, the performance improves considerably when sharing the weights between the Snake Head iterations.

V. CONCLUSION

We proposed an approach to detecting calving fronts that directly predict the desired contours instead of predicting dense masks as an intermediate output. By training our method and existing methods on the CALFIN dataset, we showed that this new approach outperforms previous methods both on the CALFIN and TUD test sets, and exhibits competitive performance on the Baumhoer test set. In our ablation study, we showed the importance of network elements like the loss function, stopping gradient flow in the Snake Head, and sharing the weights between iterations.

Furthermore, we showed that deep AC models not only provide accurate delineations of calving fronts but also are naturally suited for the quantification of the prediction uncertainties.

We hope that this study can inspire new ways of approaching similar tasks in remote sensing where boundaries are studied, like grounding line detection or firn line detection.

Finally, we believe that the shift in representation from pixel-wise masks to GIS-native data structures like polylines will not only reduce the computational burden but also allow for exciting new approaches like enforcing physical constraints and temporal consistency or analysis across different coordinate reference systems.

REFERENCES

[1] J. Mouginot et al., "Forty-six years of Greenland ice sheet mass balance from 1972 to 2018," *Proc. Nat. Acad. Sci. USA*, vol. 116, no. 19, pp. 9239–9244, May 2019.

[2] M. Tedesco, S. Doherty, X. Fettweis, P. Alexander, J. Jeyaratnam, and J. Stroeve, "The darkening of the Greenland ice sheet: Trends, drivers, and projections (1981–2100)," *Cryosphere*, vol. 10, no. 2, pp. 477–496, Mar. 2016.

[3] A. Shepherd et al., "Mass balance of the Greenland ice sheet from 1992 to 2018," *Nature*, vol. 579, no. 7798, pp. 233–239, Mar. 2020.

[4] D. Cheng et al., "Calving front machine (CALFIN): Glacial termini dataset and automated deep learning extraction method for Greenland, 1972–2019," *Cryosphere*, vol. 15, no. 3, pp. 1663–1675, 2021.

[5] K. Heidler, L. Mou, C. Baumhoer, A. Dietz, and X. X. Zhu, "HED-UNet: Combined segmentation and edge detection for monitoring the Antarctic coastline," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4300514.

[6] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *Int. J. Comput. Vis.*, vol. 1, no. 4, pp. 321–331, Jan. 1988.

[7] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. 33rd Int. Conf. Mach. Learn.*, M. F. Balcan and K. Q. Weinberger, Eds. New York, NY, USA, vol. 48, Jun. 2016, pp. 1050–1059.

[8] C. Baumhoer, A. Dietz, S. Dech, and C. Kuenzer, "Remote sensing of Antarctic glacier and ice-shelf front dynamics—A review," *Remote Sens.*, vol. 10, no. 9, p. 1445, Sep. 2018.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2015, pp. 234–241.

[10] Y. Mohajerani, M. Wood, I. Velicogna, and E. Rignot, "Detection of glacier calving margins with convolutional neural networks: A case study," *Remote Sens.*, vol. 11, no. 1, p. 74, Jan. 2019.

[11] C. A. Baumhoer, A. J. Dietz, C. Kneisel, and C. Kuenzer, "Automated extraction of Antarctic glacier and ice shelf fronts from Sentinel-1 imagery using deep learning," *Remote Sens.*, vol. 11, no. 21, p. 2529, Oct. 2019.

[12] M. Periyasamy, A. Davari, T. Seehaus, M. Braun, A. Maier, and V. Christlein, "How to get the most out of U-Net for glacier calving front segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1712–1723, 2022.

[13] E. Loebel et al., "Extracting glacier calving fronts by deep learning: The benefit of multispectral, topographic, and textural input features," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4306112.

[14] M. Holzmann, A. Davari, T. Seehaus, M. Braun, A. Maier, and V. Christlein, "Glacier calving front segmentation using attention U-Net," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2021, pp. 3483–3486.

[15] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Computer Vision—ECCV 2018*, vol. 11211. Cham, Switzerland: Springer, 2018, pp. 833–851.

[16] E. Zhang, L. Liu, L. Huang, and K. S. Ng, "An automated, generalized, deep-learning-based method for delineating the calving fronts of Greenland glaciers from multi-sensor remote sensing imagery," *Remote Sens. Environ.*, vol. 254, Mar. 2021, Art. no. 112265.

[17] N. Gourmelon, T. Seehaus, M. Braun, A. Maier, and V. Christlein, "Calving fronts and where to find them: A benchmark dataset and methodology for automatic glacier calving front extraction from synthetic aperture radar imagery," *Earth Syst. Sci. Data*, vol. 14, no. 9, pp. 4287–4313, Sep. 2022. [Online]. Available: https://essd.copernicus.org/articles/14/4287/2022/

[18] M. Marochov, C. R. Stokes, and P. E. Carbonneau, "Image classification of marine-terminating outlet glaciers in Greenland using deep learning methods," *Cryosphere*, vol. 15, no. 11, pp. 5041–5059, Nov. 2021.

[19] J. Liu, E. M. Enderlin, H.-P. Marshall, and A. Khalil, "Automated detection of marine glacier calving fronts using the 2-D wavelet transform modulus maxima segmentation method," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 11, pp. 9047–9056, Nov. 2021.

[20] A. Davari, C. Baller, T. Seehaus, M. Braun, A. Maier, and V. Christlein, "Pixelwise distance regression for glacier calving front detection and segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5224610.

[21] T. Klinger, M. Ziems, C. Heipke, H. W. Schenke, and N. Ott, "Antarctic coastline detection using snakes," *Photogramm. Fernerkund. Geoinf.*, vol. 2011, no. 6, pp. 421–434, Dec. 2011.

[22] C. Rupprecht, E. Huaroc, M. Baust, and N. Navab, "Deep active contours," 2016, *arXiv:1607.05074Cs*.

[23] S. Peng, W. Jiang, H. Pi, X. Li, H. Bao, and X. Zhou, "Deep snake for real-time instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 8530–8539.

[24] Z. Liu, J. H. Liew, X. Chen, and J. Feng, "DANCE: A deep attentive contour model for efficient instance segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Waikoloa, HI, USA, Jan. 2021, pp. 345–354.

[25] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, "The importance of skip connections in biomedical image segmentation," in *Deep Learning and Data Labeling for Medical Applications*. Cham, Switzerland: Springer, 2016, pp. 179–187.

[26] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 1800–1807.

[27] A. van den Oord et al., "Parallel WaveNet: Fast high-fidelity speech synthesis," in *Proc. 35th Int. Conf. Mach. Learn. (ICML)*, J. G. Dy and A. Krause, Eds. Stockholmsmässan, Stockholm, Sweden, vol. 80, Jul. 2018, pp. 3915–3923. [Online]. Available: http://proceedings.mlr.press/v80/oord18a.html

[28] Z. Wang and Y.-J. Liu, "Active contour model by combining edge and region information discrete dynamic systems," *Adv. Mech. Eng.*, vol. 9, no. 3, 2017, Art. no. 1687814017692947.

[29] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-26, no. 1, pp. 43–49, Feb. 1978.

[30] M. Cuturi and M. Blondel, "Soft-DTW: A differentiable loss function for time-series," in *Proc. 34th Int. Conf. Mach. Learn.*, D. Precup and Y. W. Teh, Eds. vol. 70, Aug. 2017, pp. 894–903.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–15.

[32] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *Proc. 5th Int. Conf. Learn. Represent.*, Toulon, France, Apr. 2017, pp. 1–16.

[33] J. Bradbury et al. (2018). *JAX: Composable Transformations of Python+NumPy Programs*. [Online]. Available: http://github.com/google/jax

[34] T. Hennigan, T. Cai, T. Norman, and I. Babuschkin. (2020). *Haiku: Sonnet for JAX*. [Online]. Available: http://github.com/deepmind/dm-haiku

[35] J. Avbelj, R. Müller, and R. Bamler, "A metric for polygon comparison and building extraction evaluation," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 1, pp. 170–174, Jan. 2015.

[36] H. Alt and M. Godau, "Computing the Fréchet distance between two polygonal curves," *Int. J. Comput. Geometry Appl.*, vol. 5, nos. 1–2, pp. 75–91, Mar. 1995.

[37] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th Intl. Conf. Mach. Learn.*, Sydney, NSW, Australia, 2017, pp. 1321–1330.

[38] C. Dechesne, P. Lassalle, and S. Lefèvre, "Bayesian U-Net: Estimating uncertainty in semantic segmentation of Earth observation images," *Remote Sens.*, vol. 13, no. 19, p. 3836, Sep. 2021.

[39] A. Hartmann, A. Davari, T. Seehaus, M. Braun, A. Maier, and V. Christlein, "Bayesian U-Net for segmenting glaciers in SAR imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2021, pp. 3479–3482.

Konrad Heidler (Student Member, IEEE) received the bachelor's degree in mathematics and the master's degree in mathematics in data science from the Technical University of Munich (TUM), Munich, Germany, in 2017 and 2020, respectively, where he is currently pursuing the Ph.D. degree.

His main research interests are remote sensing, computer vision, and mathematical foundations of machine learning. His research work focuses on the application of deep learning in polar regions.

Lichao Mou received the bachelor's degree in automation from the Xi'an University of Posts and Telecommunications, Xi'an, China, in 2012, the master's degree in signal and information processing from the University of Chinese Academy of Sciences (UCAS), Beijing, China, in 2015, and the Dr.-Ing. degree from the Technical University of Munich (TUM), Munich, Germany, in 2020.

He is currently a Guest Professor with the Munich AI Future Laboratory AI4EO, TUM, and the Head of the Visual Learning and Reasoning Team, Department "EO Data Science," Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Weßling, Germany. Since 2019, he has been an AI Consultant with the Helmholtz Artificial Intelligence Cooperation Unit (HAICU), Oberschleißheim, Germany. In 2015, he spent six months at the Computer Vision Group, University of Freiburg, Freiburg im Breisgau, Germany. In 2019, he was a Visiting Researcher with the Cambridge Image Analysis Group (CIA), University of Cambridge, Cambridge, U.K. From 2019 to 2020, he was a Research Scientist with DLR-IMF.

Dr. Mou was a recipient of the First Place in the 2016 IEEE Geoscience and Remote Sensing Society (GRSS) Data Fusion Contest and the Finalist for the Best Student Paper Award at the 2017 Joint Urban Remote Sensing Event and the 2019 Joint Urban Remote Sensing Event.

Erik Loebel received the bachelor's degree in geodesy and geoinformation and the master's degree in geodesy from the Technische Universität Dresden (TU Dresden), Dresden, Germany, in 2016 and 2019, respectively, where he is currently pursuing the Ph.D. degree with the Chair of Geodetic Earth System Research.

His research interests include remote sensing and Earth observation, machine learning, signal processing, and time series analysis, with a special focus on the cryosphere and polar regions. His research aims at developing and applying deep learning methods for monitoring glacier changes in Greenland.

Mirko Scheinert received the Ph.D. degree in orbit dynamics of low-flying satellites from the University of Stuttgart, Stuttgart, Germany, in 1994.

He contributes to graduate lectures in geodesy and is active in the supervision of Ph.D. students. He is a Senior Scientist with the Chair of Geodetic Earth System Research, Technische Universität Dresden (TU Dresden), Dresden, Germany. His research focuses on the determination of solid Earth deformation by geodetic methods (GNSS and gravimetry), gravity field analysis and geoid determination, the application of airborne methods in geodesy, and geodetic Earth system research. For more than 25 years, he has been active in polar research, among others conducting in situ measurements in Antarctica, Greenland, and Patagonia.

Sébastien Lefèvre (Senior Member, IEEE) received the M.Sc. degree in computer engineering from the University of Technology of Compiègne, Compiègne, France, in 1999, the Ph.D. degree in computer science from the University of Tours, Tours, France, in 2002, and the Habilitation degree in computer science from the University of Strasbourg, Strasbourg, France, in 2009.

Since 2010, he has been a Full Professor of computer science with the Université Bretagne Sud, Vannes, France, where he is involved in numerous activities related to artificial intelligence for Earth and environment observation. He is the Chair of the GeoData Science Track, EMJMD Copernicus Master in Digital Earth; the Founder of the OBELIX Team, Institute for Research in Computer Science and Random Systems; and initiated the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD) Machine Learning for Earth Observation (MACLEAN) Workshop series on machine learning for Earth observation.

Xiao Xiang Zhu (Fellow, IEEE) received the M.Sc., Dr.-Ing., and Habilitation degrees in signal processing from the Technical University of Munich (TUM), Munich, Germany, in 2008, 2011, and 2013, respectively.

She is currently the Chair Professor for data science in Earth observation with TUM and was the founding Head of the Department "EO Data Science," Remote Sensing Technology Institute, German Aerospace Center (DLR), Weßling, Germany. Since 2019, she has been a Coordinator of the Munich Data Science Research School (www.mu-ds.de). Since 2019, she has been heading the Helmholtz Artificial Intelligence—Research Field "Aeronautics, Space and Transport." Since May 2020, she has been the Principal Investigator and the Director of the International Future AI Laboratory "AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond," Munich. Since October 2020, she has been serving as the Director for the Munich Data Science Institute (MDSI), TUM. She was a Guest Scientist or a Visiting Professor with the Italian National Research Council (CNR-IREA), Naples, Italy, in 2009; Fudan University, Shanghai, China, in 2014; The University of Tokyo, Tokyo, Japan, in 2015; and the University of California at Los Angeles, Los Angeles, CA, USA, in 2016. She is currently a Visiting AI Professor with the ESA's Phi-lab. Her main research interests are remote sensing and Earth observation, signal processing, machine learning, and data science, with their applications in tackling societal grand challenges, such as global urbanization, UN's sustainable development goals (SDGs), and climate change.

Dr. Zhu is a member of the Young Academy (Junge Akademie/Junges Kolleg) at the Berlin-Brandenburg Academy of Sciences and Humanities, the German National Academy of Sciences Leopoldina, and the Bavarian Academy of Sciences and Humanities. She serves on the Scientific Advisory Board of several research organizations, including the German Research Center for Geosciences (GFZ) and the Potsdam Institute for Climate Impact Research (PIK). She is also an Associate Editor of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and an Area Editor for Special Issues of *IEEE Signal Processing Magazine*.

A.3 A High-Resolution Calving Front Data Product for Marine-Terminating Glaciers in Svalbard

Reference

T. Li, K. Heidler, L. Mou, Á. Ignéczi, X. X. Zhu, and J. L. Bamber, "A high-resolution calving front data product for marine-terminating glaciers in Svalbard," *Earth System Science Data*, vol. 16, no. 2, pp. 919–939, 2024. DOI: 10.5194/essd-16-919-2024

Copyright

Earth Syst. Sci. Data, 16, 919–939, 2024
https://doi.org/10.5194/essd-16-919-2024

Open Access Earth System Science Data

A high-resolution calving front data product for marine-terminating glaciers in Svalbard

Tian Li[1,2], Konrad Heidler[2], Lichao Mou[2], Ádám Ignéczi[1], Xiao Xiang Zhu[2,3], and Jonathan L. Bamber[1,2]

[1]Bristol Glaciology Centre, School of Geographical Sciences, University of Bristol, Bristol BS8 1SS, UK
[2]Chair of Data Science in Earth Observation, Department of Aerospace and Geodesy, Technical University of Munich, Munich 80333, Germany
[3]Munich Center for Machine Learning, Technical University of Munich, Munich 80333, Germany

Correspondence: Tian Li (tian.li@bristol.ac.uk)

Received: 28 September 2023 – Discussion started: 16 October 2023
Revised: 20 December 2023 – Accepted: 24 December 2023 – Published: 20 February 2024

Abstract. The mass loss of glaciers outside the polar ice sheets has been accelerating during the past several decades and has been contributing to global sea-level rise. However, many of the mechanisms of this mass loss process are not well understood, especially the calving dynamics of marine-terminating glaciers, in part due to a lack of high-resolution calving front observations. Svalbard is an ideal site to study the climate sensitivity of glaciers as it is a region that has been undergoing amplified climate variability in both space and time compared to the global mean. Here we present a new high-resolution calving front dataset of 149 marine-terminating glaciers in Svalbard, comprising 124 919 glacier calving front positions during the period 1985–2023 (https://doi.org/10.5281/zenodo.10407266, Li et al., 2023). This dataset was generated using a novel automated deep-learning framework and multiple optical and SAR satellite images from Landsat, Terra-ASTER, Sentinel-2, and Sentinel-1 satellite missions. The overall calving front mapping uncertainty across Svalbard is 31 m. The newly derived calving front dataset agrees well with recent decadal calving front observations between 2000 and 2020 (Kochtitzky and Copland, 2022) and an annual calving front dataset between 2008 and 2022 (Moholdt et al., 2022). The calving fronts between our product and the latter deviate by 32 ± 65 m on average. The R^2 of the glacier calving front change rates between these two products is 0.98, indicating an excellent match. Using this new calving front dataset, we identified widespread calving front retreats during the past four decades, across most regions in Svalbard except for a handful of glaciers draining the ice caps Vestfonna and Austfonna on Nordaustlandet. In addition, we identified complex patterns of glacier surging events overlaid with seasonal calving cycles. These data and findings provide insights into understanding glacier calving mechanisms and drivers. This new dataset can help improve estimates of glacier frontal ablation as a component of the integrated mass balance of marine-terminating glaciers.

1 Introduction

Glaciers and ice caps (GIC) distinct from the Greenland and Antarctic ice sheets are a significant contributor to global sea-level rise in addition to thermal expansion (Intergovernmental Panel on Climate Change, 2023; Meredith et al., 2019). Their mass loss has been accelerating during the early twenty-first century and their thinning rates have doubled (Hugonnet et al., 2021). Specifically, the mass loss from Arctic glaciers during 2006–2015 contributed to sea-level rise at a similar rate (0.6 ± 0.1 mm yr^{-1}) to the Greenland Ice Sheet in response to the accelerated warming trend in the Arctic (Intergovernmental Panel on Climate Change, 2023). Recent observations show that the maximum warming rate on Earth (> 1.25 °C per 10 years) during 1979–2021 lies in the Russian Arctic close to Svalbard (Rantanen et al., 2022), which

is one of the most climatically sensitive regions in the world (van Pelt et al., 2018; Serreze and Barry, 2011).

Svalbard is an Arctic Archipelago located near the northeast coast of Greenland and lies close to the northern limit of warm North Atlantic water (Nuth et al., 2010). Its climate displays extreme variability in both space and time. The southwest region has milder and more humid conditions while the northeast is colder and drier (Schuler et al., 2020), making it an ideal region for studying the response of glaciers to climatic forcing. In Svalbard, the warming rate has been 1.7 °C per 10 years since 1991, about 7 times the global average (Nordli et al., 2020). Glaciers on Svalbard have been losing mass since the 1960s with a trend towards a more negative mass balance since 2000 (Schuler et al., 2020; Nuth et al., 2010). High-resolution regional climate models reveal that modest atmospheric warming in the mid-1980s forced the limit of the firn zone (the boundary between ice and compacted snow) to the hypsometric peak, leading to firn cover reduction, albedo reduction, and increased surface runoff, amplifying the mass loss from all elevations (Noël et al., 2020). By linking historical and modern glacier observations, it was predicted that the twenty-first-century glacier thinning rates in Svalbard would be more than double the rates of 1936–2010, with a strong dependence on air temperature (Geyman et al., 2022).

Despite recent progress in estimating the mass balance of glaciers in Svalbard, uncertainties remain, especially the quantification of frontal ablation – a combination of calving and basal melting. Frontal ablation is a key component of the total mass balance of marine-terminating glaciers, with the other being the climatic mass balance (Schuler et al., 2020). Despite its importance, most global glacier models do not include the frontal ablation component at all (Rounce et al., 2023). In Svalbard, 15 % of the glaciers are marine-terminating and in other Arctic sectors it is significantly higher (Oppenheimer et al., 2019). They account for about 60 % of the total glacierized area (Błaszczyk et al., 2009) and experienced one of the highest frontal ablation rates in the Northern Hemisphere. However, there have only been two systematic studies estimating the frontal ablation of glaciers in Svalbard (Błaszczyk et al., 2009; Kochtitzky et al., 2022). Błaszczyk et al. (2009) estimated the frontal ablation rates of 163 Svalbard tidewater glaciers during a short period from 2000 to 2006. Kochtitzky et al. (2022) updated this record by estimating the frontal ablation with a decadal time resolution for 2000–2010 and 2010–2020.

One major limitation of frontal ablation estimates is the scarcity in calving front observations of marine-terminating glaciers (Kochtitzky et al., 2023), which is essential for determining the relative contributions of calving and submarine melting (Schuler et al., 2020) and their governing processes. A detailed understanding of the calving mechanism and its drivers is crucial for the accurate prediction of glacier response to future climate forcing and consequent sea-level change (Benn et al., 2007; Kochtitzky et al., 2023). The currently available calving front datasets for marine-terminating glaciers in Svalbard are limited to either a small sample of glaciers (Murray et al., 2015; Strozzi et al., 2017; Holmes et al., 2019; Nuth et al., 2019) or to low temporal resolutions in calving front observations (Błaszczyk et al., 2009; Carr et al., 2017; Kochtitzky and Copland, 2022; Nuth et al., 2013; Moholdt et al., 2022).

Calving front mapping of glaciers beyond the Greenland Ice Sheet has primarily relied on manual delineation from optical satellite imagery such as Landsat and ASTER (McNabb and Hock, 2014; Kochtitzky and Copland, 2022; Cook et al., 2019). This often results in low spatial coverage and temporal resolution, as optical images are often influenced by the presence of clouds and the polar night. With the availability of new optical satellite missions such as Sentinel-2 and Landsat-9, as well as the SAR satellite Sentinel-1, it is possible to achieve a short image acquisition interval of 1–3 d all year round. In the meantime, the growing availability of extensive satellite catalogues imposes a challenge for manual delineation. There is, therefore, a need for efficient automated methods. In recent years, deep learning has demonstrated promising capabilities in accurately mapping glacier calving fronts (Mohajerani et al., 2019a; Cheng et al., 2021; Heidler et al., 2022; Loebel et al., 2023; Gourmelon et al., 2022; Zhang et al., 2019; Baumhoer et al., 2019, 2023). Mohajerani et al. (2019) pioneered the application of deep learning in glacier calving front mapping by developing a U-Net architecture to isolate the calving front from satellite images. The method was tested on Helheim Glacier in Greenland with a mean deviation of 96.3 m from ground truth, which is a manually mapped calving front from Landsat images. Building on this, Heidler et al. (2022) proposed a novel deep-learning framework, HED-UNet, by combing semantic segmentation and edge detection, which outperforms the traditional U-Net framework. So far, these deep-learning frameworks have only been applied to a small sample of glaciers mainly located on the Greenland and Antarctic ice sheets. Nonetheless, these case studies serve as a foundation for automated, high-temporal-resolution mapping of glacier terminus locations on a large spatial scale for glaciers outside the ice sheets.

Here, we introduce a novel automated processing pipeline designed to map glacier calving fronts using a new deep-learning framework Charting Outlines by Recurrent Adaptation (COBRA), which outperforms image segmentation models by combining convolutional neural networks and active contour models for calving front mapping (Heidler et al., 2023). This study yields a new high-resolution glacier calving front data product containing 124 919 calving front traces for 149 marine-terminating glaciers in Svalbard during the period 1985–2023 (Li et al., 2023), utilizing data from multiple optical and SAR satellite sensors, including Landsat, ASTER, Sentinel-2, and Sentinel-1. This newly compiled dataset offers unprecedented temporal density, which is valuable for analysing both the seasonal and interannual

variations in glacier calving fronts, as well as capturing surge events.

2 Data and methodology

2.1 Automated satellite image downloading from Google Earth Engine

To generate the calving front data product, optical images from three different satellite platforms – Landsat, Terra-ASTER, and Sentinel-2 – along with SAR images in the Extra Wide (EW) swath mode from Sentinel-1, spanning the period from 1972 to January 2023, were used. The reason for using the EW mode of Sentinel-1 images instead of the higher-resolution Interferometric Wide (IW) mode is that the EW mode has greater coverage over Svalbard. The satellite images were acquired from the Google Earth Engine (GEE) platform with a diverse range of image resolutions, repeat cycles, and operation durations shown in Table 1. The detailed workflow for downloading satellite images automatically for marine-terminating glaciers from different GEE satellite image collections (Table A1 in the Appendix) is shown in Fig. 1.

Our selection of glaciers in Svalbard is based on the tidewater glacier terminus data product generated by Kochtitzky and Copland (2022) which includes areal change polygons for all marine-terminating glaciers across the Arctic in two different periods: 2000–2010 and 2010–2020. To begin, the Kochtitzky and Copland (2022) frontal areal change polygons of each glacier were used to produce the glacier domain shapefiles (Box 1 in Fig. 1). For each glacier, all the available different areal change polygons generated in two different time periods were first merged into one single polygon. Then the minimum bounding rectangle (MBR) of this merged polygon was generated. The final glacier domain polygon (black boxes in Fig. 2a) was produced by adding a 1500 m buffer length to the MBR. If the final glacier domain polygon contained multiple polygons likely to be associated with tributary glaciers, these polygons were then divided into separate individual glacier area change polygons and assigned unique identifiers by adding sequential letters to its original Randolph Glacier Inventory (RGI) version 6 glacier id (RGI Consortium, 2017) as a new glacier id; this updated glacier id was used throughout the study. In total, we generated 220 glacier domain shapefiles (hereinafter referred to as 220 marine-terminating glaciers – we took tributary glacier as an independent glacier) (black boxes in Fig. 2). The domain shapefile was used in defining the glacier spatial extent to be used in querying satellite images from the GEE API.

For each glacier domain, satellite images were retrieved from four distinct satellite platforms, namely Landsat 1-9, Terra-ASTER, Sentinel-2A/B, and Sentinel-1A/B (Table 1). The images were downloaded throughout the entire time span of each satellite mission and were used in mapping the

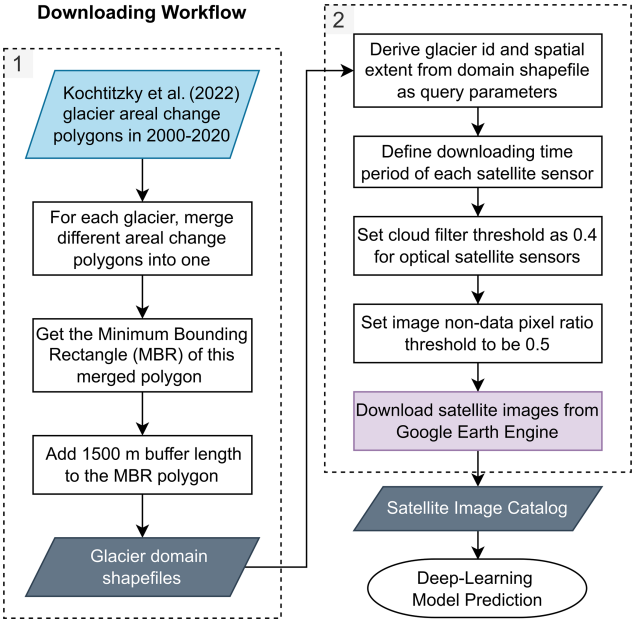


Figure 1. The workflow of generating glacier domain shapefiles (box 1) and automated downloading satellite images from Google Earth Engine (GEE) (box 2) for Svalbard marine-terminating glaciers. The coloured geometries indicate key inputs and outputs.

glacier calving front locations. For optical satellite images downloaded from Landsat, Terra-ASTER, and Sentinel-2, we set a cloud filter threshold of 40 %. Furthermore, a universal threshold for a non-data pixel ratio per image is set as 50 % for both optical and SAR images. If the proportion of non-data pixels in a given satellite image exceeds 50 %, it is presumed that this image may lack a sufficient number of pixels for accurate predictions. In addition, we did not merge satellite images acquired on the same day considering the large number of data available. For the 220 marine-terminating glacier domains in Svalbard, 1 135 074 satellite images were downloaded for the glacier calving front prediction over the period 1972–2023 in our study.

2.2 Deep-learning model and pre-processing

We used the deep-learning model Charting Outlines by Recurrent Adaptation (COBRA) to predict the glacier calving front locations. The COBRA model combines a convolutional neural network (CNN) for feature extraction and an active contour model for the delineation (Heidler et al., 2023). Unlike the traditional image segmentation models such as CALFIN (Cheng et al., 2021) and HED-UNet (Heidler et al., 2022) which separate an image into land-ice and ocean classes, the COBRA model can directly output the calving front line segment as a shapefile instead of recovering the vectorized contour from intermediate predictions in a semantic segmentation approach. Figure 3a shows the model architecture, and it comprises two different components: a back-

Table 1. Image resolutions of different satellite sensors used in the calving front mapping.

Satellite platform	Resolution	Availability	Repeat cycle	Band
Landsat	30 m	1972	16 d	Near-infrared band
ASTER	30 m	2000	16 d	Near-infrared band
Sentinel-2	10 m	2015	10 d	Near-infrared band
Sentinel-1	40 m	2014	12 d	HH band (EW mode)

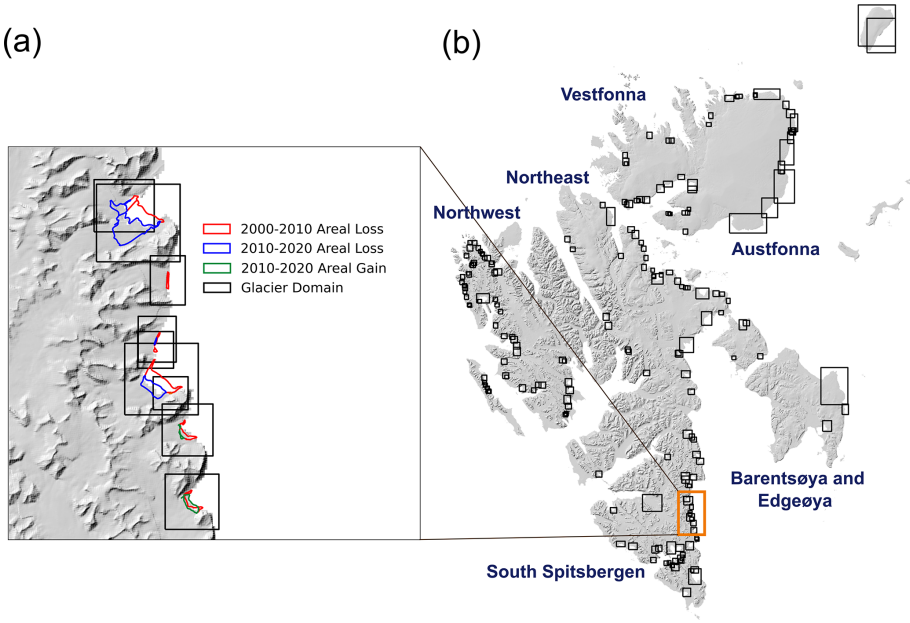


Figure 2. **(a)** Examples of glacier areal change polygons (coloured outlines) generated in Kochtitzky and Copland (2022) and the glacier domain polygons derived in this study (black boxes). The glacier areal loss during 2000–2010 is denoted as a red polygon, the glacier areal loss during 2010–2020 is denoted as a blue polygon, and the glacier areal gain during 2010–2020 is denoted as a green polygon. **(b)** The spatial distributions of 220 glacier domains generated in this study (black boxes); the orange box denotes the zoomed-in region shown in panel **(a)**. The background hillshade map is generated from the 50 m resolution Svalbard digital elevation model (DEM) (https://data.npolar.no/dataset/dce53a47-c726-4845-85c3-a65b46fe2fea, last access: 18 April 2023).

bone and a prediction head. The backbone of the COBRA architecture utilizes a versatile two-dimensional CNN to extract meaningful semantic features from the input imagery; here the Xception backbone was employed (Chollet, 2016; Cheng et al., 2021). The second component consists of a prediction head known as the "snake head", which leverages the feature map of the backbone to generate the ultimate network predictions. The snake head starts with an initial calving front contour with vertices generated in the centre of the image, then progressively refines the contour by incorporating sampled values from the feature map extracted from the backbone network and iterating this process four times (Heidler et al., 2023). The loss function of the COBRA model is based on the dynamic time warping (DTW) loss, which measures the similarity between the predicted contour and the true contour (Heidler et al., 2023). The loss function is shown as Eq. (1):

$$\mathcal{L}_{\mathrm{DTW}}(p,t) = \min_{(i_k,j_k)_{k\in[K]}\in\kappa} \sum_k ||p_{ik} - t_{ik}||_2^2, \quad (1)$$

where the predicted contour p is represented by vertices p_i with $1 \le i \le V$, and the true contour t is represented by vertices t_j with $1 \le j \le V$. κ denotes the set of all possible realignments $(i_k, j_k)_{k\in[K]}$ that satisfy the following three conditions: (1) for any $i \in \{1, \ldots V\}$ there is a k with $i_k = i$; (2) for any $j \in \{1, \ldots V\}$ there is a k with $j_k = j$; and (3) the sequences i_k and j_k are non-decreasing in k.

The model was trained for 500 epochs on the CALFIN training dataset (Cheng et al., 2021) which includes 1541 Landsat optical images and 232 Sentinel-1 SAR images for 66 Greenlandic glaciers during 1972–2019. In addition, it was tested on three different test sets including the CALFIN test set, the TU Dresden (TUD) (Loebel et al., 2022), which includes 1127 Landsat optical images in 2013–2021 for 23

glaciers, as well as the Baumhoer dataset (Baumhoer et al., 2019), which includes 62 Sentinel-1 SAR images for glaciers located in Antarctica. The Adam optimizer (Kingma and Ba, 2014) with an initial learning rate of 10^{-3} was used in the training process. The model was implemented in JAX using the Haiku framework (Heidler et al., 2023).

At the time of the COBRA model development, CALFIN was the most complete glacier calving front mapping training dataset available for the Northern Hemisphere (Cheng et al., 2021). Although the sizes of tidewater glaciers in Greenland are typically much larger than those in Svalbard, their geomorphological characteristics are similar (Benn et al., 2007). Therefore, we used this pre-trained COBRA model to map glacier calving fronts in Svalbard. In order to maintain the consistency with the training dataset (Cheng et al., 2021), the near-infrared band of the optical images and the HH band of the SAR images were used (Table 1). Each satellite image was initially cropped into a square shape, with the side length equal to the shortest dimension of the original image, centred around its midpoint. Then a min–max image scaling was applied to the cropped satellite image prior to calving front prediction. The COBRA model predicts the entire coastline including both the fjord boundary (green line in Fig. 3b) and the glacier calving front (red line in Fig. 3b) (Heidler et al., 2023); an example is shown in Fig. 3b. Therefore, the model outputs need to be post-processed to isolate the actual calving front.

2.3 Post-processing

While deep-learning techniques have demonstrated effectiveness in delineating glacier calving front locations (Cheng et al., 2021; Zhang et al., 2019; Heidler et al., 2022), many have only been trained on limited datasets, potentially missing some glacier terminus conditions in different satellite images. Consequently, due to the well-known distributional shift, the network may produce inaccurate predictions when processing satellite images that are not well-represented in the training datasets, e.g. where the calving front is less distinct, shadowing occurs, fast-ice is present or other factors. These inaccurate predictions need to be removed from the final glacier calving front data product. In addition, the COBRA model prediction includes not only the glacier calving front, but also the neighbouring fjord boundary which is not needed. Here we developed an automated post-processing pipeline to eliminate these inaccurate terminus traces and mask out the fjord boundary (Fig. 4).

The pipeline consists of four major steps: (1) preliminary filtering of the initial COBRA model outputs based on the length and curvature of calving front line segments (Box 1 in Fig. 4); (2) use of a fjord mask to exclude the fjord boundary or the other non-calving-front features of each glacier (Box 2 in Fig. 4); (3) identification and removal of erroneous traces based on glacier calving front line segment density and similarity (Box 3 in Fig. 4); (4) utilizing a predefined glacier centreline to generate a time series of calving front changes and identifying outliers by applying a median filter to the times series of the calving front change (Box 4 in Fig. 4).

2.3.1 Filter original model output based on length and curvature

In cases where the glacier calving front is heavily obscured by cloud cover or high sea-ice concentration, the calving front may be less distinguishable in satellite images and the COBRA model can generate inaccurate predictions. These can manifest as either excessively short or long line segments and can exhibit overly complicated curvature shape. The first step of the post-processing pipeline is to remove these inaccurate predictions according to the line segment length and curvature complexity (Box 1 in Fig. 4). The terminus length and curvature filtering thresholds are based on the automatic screening module developed by Zhang et al. (2023). Two thresholds T_{L} and T_{U} based on the inter-quartile range were used for all the initial terminus trace outputs from COBRA in each glacier domain:

$$T_{\mathrm{L}} = Q1 - 1.5 \times (Q3 - Q1) \quad (2)$$

$$T_{\mathrm{U}} = Q3 + 1.5 \times (Q3 - Q1), \quad (3)$$

where $Q3$ is the 75th percentile and $Q1$ is the 25th percentile of the data range. For the terminus length, we defined the terminus traces from both the lower and upper thresholds T_{L} and T_{U} as outliers because the terminus traces that are either too short or too long are likely to be anomalies. Following the length filtering of the terminus traces, we calculated the curvature of each terminus trace as the average for the curvatures between two adjacent points along each terminus trace, then eliminated the terminus traces with curvature values exceeding the upper threshold T_{U}. The reason for only applying an upper threshold for curvature complexity is because the high-quality terminus trace should be smooth with minimal curvature (Zhang et al., 2023).

2.3.2 Crop and filter glacier calving front using fjord mask

Following the initial filtering of terminus trace outputs based on the line segment length and curvature complexity in Sect. 2.3.1, a fjord mask was implemented for each glacier (yellow polygon in Fig. 3b). As the model output includes both the fjord boundary (i.e. land–water contact) and the glacier calving fronts, the fjord mask serves to exclude the fjord boundary, retaining only the calving front line segment that we are interested in (Box 2 in Fig. 4). The fjord mask was generated by combining the ice-free zone from a binary ice mask and the land zone from a binary land mask.

The binary land mask was created using the high-resolution (3"; ∼ 90 m) Water Body Mask (WBM) product – showing inland water bodies and oceans – that is supplied with the Copernicus GLO-90 digital elevation model

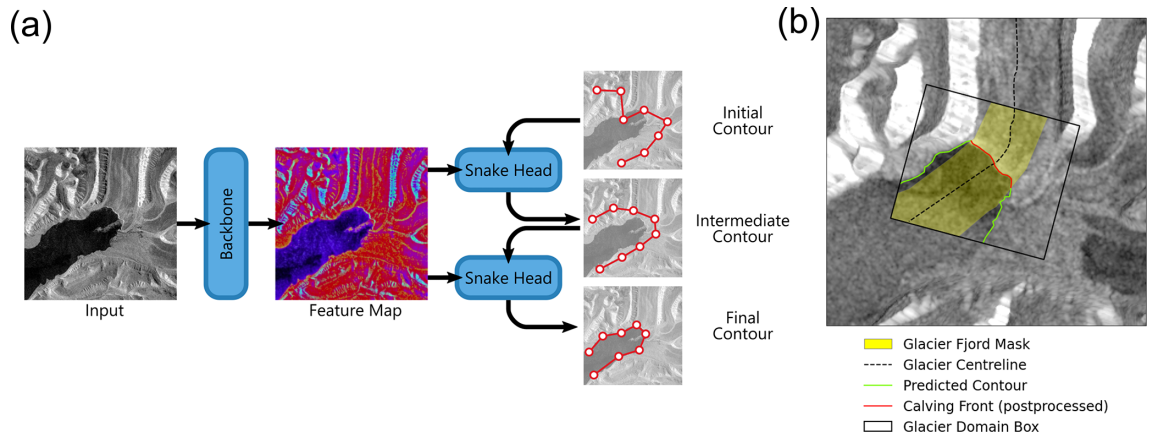


Figure 3. **(a)** The Charting Outlines by Recurrent Adaptation (COBRA) deep-learning model architecture used in this study (Heidler et al., 2023); here only two iterations of the snake head are shown. **(b)** The calving front predicted by the COBRA model from Sentinel-1A SAR image on 21 December 2022 for Tunabreen glacier (RGI60-07.01458); the glacier fjord mask is shown as a yellow polygon, the glacier centreline is shown as a dashed black line, the model output is shown as a combination of a green and red line, the post-processed final calving front is shown as a red line, and the glacier domain box is shown as a black outline.

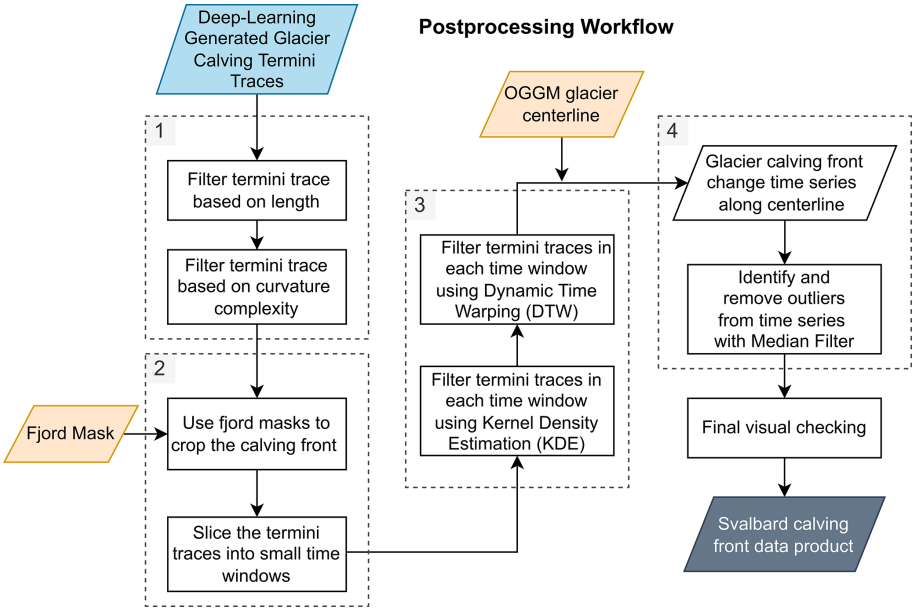


Figure 4. The flowchart of post-processing workflow applied to the glacier calving front traces mapped from the pre-trained COBRA deep-learning model. The coloured geometries indicate key inputs and outputs.

(DEM) dataset (ESA, 2021). The WBM, together with the DEM product, is referenced on the WGS-84 ellipsoid and is provided in 1° × 1° tiles globally. We used the RGI version 6 (RGI Consortium, 2017) first-order region shapefile for Svalbard to compile the appropriate list of WBM tiles. After mosaicking all the WBM tiles for Svalbard, we converted the original WBM product to a binary land mask by recategorizing all non-ocean pixels as land. The land mask mosaic was then re-projected to a 250 m grid (EPSG:3574) and clipped with the RGI region outlines. The binary ice mask was created using RGI version 6 glacier outlines for Svalbard. These are provided in shapefiles and then rasterized to the 250 m resolution land mask mosaic grid, which was applied to correct for any potential mismatches (i.e. masking out the ocean) between the RGI and Copernicus datasets.

After compiling the binary land and ice masks, we combined the two to find ice free land and vectorized the resulting product. As a final step we added a buffer zone of 200 m length to the merged ice-free land polygon then removed this buffered polygon from the glacier domain box to get the fjord mask that was used in subsequent steps (yellow polygon in Fig. 3b). All the fjord masks for our glacier do-

mains were visually checked and manually adjusted if necessary to make sure the mask can cover the entire calving front changes. The glacier calving fronts that have been clipped using fjord masks were subsequently categorized into individual time windows which were defined based on the observation density. During the period 1970–2015 when the data collection was limited, we set five distinct time windows: 1 January 1970 to 1 January 1990; 1 January 1990 to 1 January 2000; 1 January 2000 to 1 January 2005; 1 January 2005 to 1 January 2010; and 1 January 2010 to 1 January 2015. From January 2015 to January 2023, we set 17 time intervals, each spanning 6 months.

We implemented length and curvature filters as described in Sect. 2.3.1 prior to clipping calving fronts with fjord masks to avoid inaccuracies. If calving fronts are clipped first, it could result in traces with unrealistic lengths or high curvature complexity still retained within the fjord masks. Consequently, this could lead to the erroneous exclusion of high-quality calving front traces by the length and curvature filters, particularly if most of the clipped front traces are of poor quality. This is especially problematic for smaller glaciers with complex surface features that are not well represented in the CALFIN training dataset.

2.3.3 Terminus filtering based on the line segment density and similarity

Within a given time window defined in Sect. 2.3.2, we assume that the contour shapes of the majority of terminus traces are similar, and any erroneous terminus trace will significantly deviate from this expected similarity. Guided by this principle, we subsequently implemented two additional filtering steps for the clipped glacier calving front line segments: kernel density estimation (KDE) and dynamic time warping (DTW) (Box 3 in Fig. 4). KDE is a well-established nonparametric approach to estimate the continuous density function based on a sample dataset and can cope with an inhomogeneous distribution of observations (Davies et al., 2018). Here it was used to estimate the density distribution of the glacier calving fronts. We first converted all the terminus trace line segments in one glacier domain into scattered points, then calculated their kernel-density estimates using a Gaussian kernel. For the density map, we set the upper threshold as 75 % percentile $Q3$ and extracted the contour boundary of the area where the KDE density is higher than $Q3$ – the area inside this contour was taken as the boundary where glacier calving fronts are mostly likely to locate. For every terminus trace line segment, we calculated its intersection with this $Q3$ contour. Terminus traces situated completely outside the threshold contour were identified as outliers and subsequently excluded from the data product. Terminus traces completely enclosed within the threshold contour, or those >95 % of the total trace length within the contour polygon, were taken as potential valid results and retained for subsequent post-processing steps.

DTW is a technique that has been used in time series analysis to measure similarity between two sequences that vary in time and speed, and to find the optimal alignment by accommodating time shifts and local shape distortions (Müller, 2007). Here we use DTW to measure the similarity between two different terminus trace line segments. For each terminus line segment, the DTW distances between this line segment and all the remaining terminus line segments were calculated. The resulting mean value was taken as the ultimate DTW distance for this terminus trace. After iterating this step for all the terminus traces within a given time window, an outlier detection threshold of 75 % percentile $Q3$ of all the DTW distances was applied to identify the anomalous terminus traces. If the DTW distance of a given terminus trace exceeds this threshold, it was eliminated from subsequent processing.

2.3.4 Calving front change time series and median filtering

The primary objective of measuring glacier calving front locations is to determine changes over time. Therefore, as a final step, we generated a time series of the calving front change for each glacier using a centreline approach and used this to remove outliers. The centreline approach measures the advance or retreat of the glacier calving front along a glacier centreline in relation to their earliest position (Cheng et al., 2021). The glacier centrelines for all the marine-terminating glaciers analysed in this study were first derived using the Open Global Glacier Model (OGGM) (Maussion et al., 2019). The OGGM glacier centreline was based on a predefined glacier domain boundary from the RGI glacier database (Pfeffer et al., 2014), and therefore its length may not cover all the calving front traces mapped in this study as some glaciers undergo dramatic changes at their calving fronts during the study period. To address this issue, we automatically extended the endpoint of each OGGM centreline by an additional 10 km in the seaward direction, following the direction defined by the line segment connecting the two outermost seaward data points of the OGGM centreline. In addition, only the main glacier centreline was extracted from the OGGM model; for glacier domains located at the tributary glaciers we manually mapped the glacier centrelines. All the glacier centrelines were visually checked and modified when necessary to make sure it covers the entire glacier calving front locations of a given glacier and is near perpendicular to the calving front.

To make use of the dense glacier calving front observations after 2014, a rolling window of 10 observations was applied. Note we did not apply a rolling window for observations prior to 2014 due to the lack of sufficient terminus traces because the available trace number within 1 year could be less than 10 (Fig. A1). We first calculated an upper threshold as the greater value between 200 m and the maximum standard deviation of calving front changes in all rolling win-

dows. The range between the median calving front change distance in each rolling window above and below this threshold serves as the criterion for identifying and removing outliers. This assumes that within a short period of time with 10 observations, the glacier calving front change distance is likely to be less than 200 m (Luckman et al., 2015). Furthermore, utilizing the highest standard deviation of calving front change observed across all rolling windows could accommodate the occurrence of large calving events. Although this threshold may not sufficiently capture all the large calving events which are mostly stochastic events that are difficult to detect automatically, the calving of large tabular icebergs is less likely to happen in Svalbard. Nonetheless, this criterion will need to be further improved for large tidewater glaciers in the Greenland Ice Sheet.

As a final step, all the glacier terminus traces after the above post-processing steps were visually checked to make sure they are correct. The examples of different post-processed glacier calving front traces for four different satellite sensors under different environmental conditions are shown as solid red lines in Fig. 5. In total, 206 371 glacier calving fronts were identified by the automated post-processing steps and 81 452 terminus traces were discarded in the visual checking. The ratio of successful calving front delineations (124 919) compared to all the input satellite images (1 135 074) is 12 %. The high abandonment rate could be attributed to three factors: (1) some satellite images may not fully capture the glacier calving front, as we did not merge the same-day images, preventing successful delineation; (2) our post-processing workflow uses multiple interquartile range filters across different steps, which can significantly reduce the output quantity; and (3) the extensive satellite images downloaded from GEE permit a strict post-processing regime, and this can improve our confidence in calving front delineation and minimizing manual checks, given that COBRA was trained on a limited training dataset from Greenland tidewater glaciers.

3 Results

3.1 Dataset overview

Using the methodology developed in this study, we produced a new high-resolution calving front dataset which contains 124 919 glacier calving fronts for 149 marine-terminating glaciers (based on updated glacier ids in Sect. 2.1) in Svalbard over the period 1985–2023 (Li et al., 2023). The final product includes only 149 glaciers, fewer than the 220 glacier domains used, because glaciers that became land-terminating during the study period were excluded, and glaciers that had too few calving fronts due to lack of satellite images were discarded in the rigorous post-processing steps. The dataset is presented as a single GeoPackage file containing five different layers: glacier domains generated in Sect. 2.1, fjord masks generated in Sect. 2.3.2, glacier centrelines generated in Sect. 2.3.4, glacier calving front terminus traces mapped in this study, and the along-centreline glacier calving front change time series in relation to the earliest time stamp. Each layer contains 149 different geometry features representing 149 marine-terminating glaciers. The detailed metadata provided in this GeoPackage file are shown in Table 2, including information on glacier id, satellite platform, satellite image id, satellite image acquisition date, and the glacier calving front change distance along the centreline. In addition, we also provided spatial distribution map plots of the glacier calving front traces and line plots depicting the time series of calving front changes for each individual glacier. These plots are provided in PNG file format and can be accessed in the figures folder.

The greatest number of traces was obtained after 2014 due to the availability of Sentinel-1 and Sentinel-2 satellites (Figs. 6, 7, and A1); the low trace number in 2023 is because we only downloaded images in January. The annual average number of traces per glacier between 2014 and 2022 is 100, representing an average temporal resolution of 4 d. This allows us to discern the seasonal patterns of glacier calving front changes. We demonstrate this in the case of five glaciers across Svalbard, including a surging glacier Osbornebreen, that exhibit strong seasonal signals after 2014 (Fig. 7). A glacier's calving fronts normally retreat (upward trend in time series) during the Arctic summer and autumn, and readvance (downward trend in time series) during the Arctic winter and spring. The manually mapped areal change polygons of Kochtitzky and Copland (2022) only contain three different calving front traces for the years 2000, 2010, and 2020; thus, this dataset cannot resolve any seasonal cycles or sudden changes in glacier calving front locations such as the surging event shown in Fig. 7l. However, these polygons align well with our calving front traces (Fig. 7b, e, h, k, n).

3.2 Uncertainty and validation

3.2.1 Uncertainty measurement

The accuracy of the predicted calving front locations from the COBRA deep-learning model depends on the spatial resolution of satellite images, the presence of cloud and shadow in optical images, speckle noises in SAR images, and the local sea-ice conditions in front of the glacier terminus. The uncertainties related to the COBRA model have been evaluated by cross-validation on three different test datasets and by comparing with different deep-learning models that were trained on the same training datasets; details can be found in Heidler et al. (2023). The average prediction error of COBRA on the CALFIN test set is 99 ± 10 m, while it is 99 ± 12 m for the Baumhoer dataset (Heidler et al., 2023). The rigorous post-processing steps developed in Sect. 2 were able to eliminate the erroneous terminus trace predictions effectively. However, the measurement error still remains even

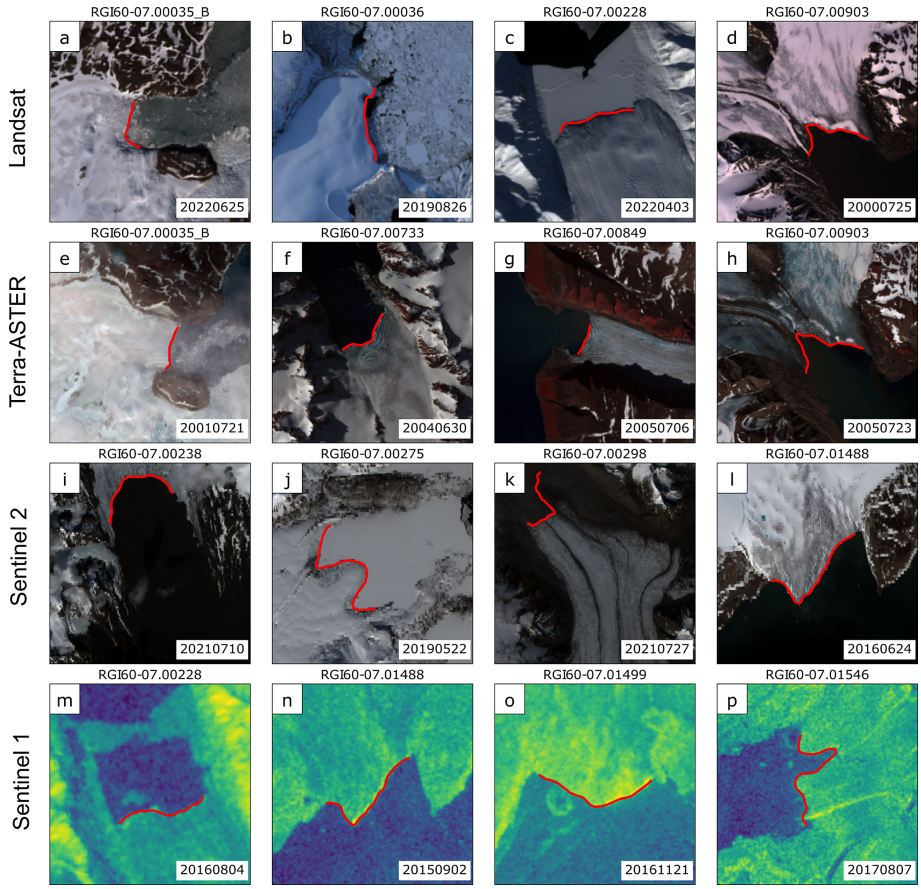


Figure 5. Examples of the post-processed glacier calving front traces for different satellite images from four satellite platforms including Landsat **(a–d)**, Terra-ASTER **(e–h)**, Sentinel-2 **(i–l)**, and Sentinel-1 **(m–p)**. The solid red lines are the final glacier terminus traces after post-processing.

Table 2. Glacier calving front trace metadata recorded in the data product.

Data field	Description
Glacier	The Randolph Glacier Inventory (RGI) version 6 glacier id
Sensor	The satellite platform used in mapping glacier calving front, including "Landsat", "Terra-ASTER", "Sentinel2" and "Sentinel1"
ImageId	The image id of the satellite image used in mapping the glacier calving front
DateString	The datetime string of the satellite image in the format of "YYYYMMDD"
CFL_Change	The calving front location (CFL) change in metres along the glacier centreline in relation to the earliest calving front location in the time series

after post-processing and varies with different satellite images obtained at different times as the environmental conditions at the glacier calving front are different. To estimate the calving front mapping uncertainty in our final data product, we compare different terminus traces mapped on the same day for a given glacier by measuring the mean distance error in their calving front locations, which is calculated as the area between two curves normalized by the average length of the curves (Cheng et al., 2021; Loebel et al., 2023). The average mean distance error in days with multiple traces is then taken as the calving front mapping uncertainty of this glacier (Fig. 8a). This is based on the hypothesis that calving front remains unchanged over a 24 h period, and traces generated from different images during the same day should be the same. Mean distance error utilizes the entire calving front trace, and therefore the estimated uncertainty is insen-

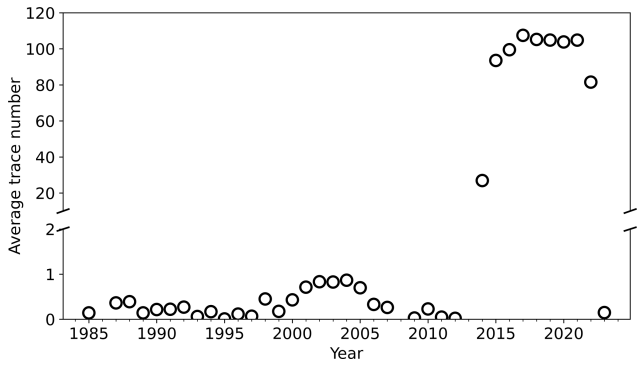


Figure 6. Average calving front traces for all marine-terminating glaciers analysed in this study from 1985 to January 2023 (also see Fig. A1 for detailed trace number of each glacier).

sitive to the centreline location and is representative for the glaciers analysed in our study. The total number of terminus traces obtained on the same day in our data product is 17 106. They span the entire time series but temporally cluster in the period 2013–2022 (Fig. 8b). Nonetheless, 88 % of the evaluated glaciers have uncertainty less than 50 m (Fig. 8c). On average, the mean distance error across Svalbard is 31 ± 30 m.

3.2.2 Validation with another data product

To further assess the glacier calving front dataset produced in this study, we calculated the mean distance error by comparing it with the Moholdt et al. (2022) annual glacier calving front data product as part of the Copernicus Glacier Service project. This product is the most complete glacier calving front data product for Svalbard prior to our study. It contains 12 years of calving front traces between 2008 and 2022 for 202 marine-terminating glaciers, and the total number of glacier terminus traces is 2419 (Table 3). Moholdt et al. (2022) generated annual shapefiles of the marine-terminating glacier calving fronts by manual delineation from optical satellite imagery mainly available from Landsat-8 and Sentinel-2 during the period 15 August–15 September of each year. Using the same approach as in Sect. 3.2.1, we calculated the mean distance error of terminus traces mapped on the same day across these two different datasets for a given glacier; the average mean distance error in days with multiple traces is then taken as the calving front mapping uncertainty of this glacier (Fig. 9a). Since the spatial coverages of the terminus traces mapped on the same day between these two data products may be significantly different, a direct comparison can result in an excessively large areal change as well as the mean distance error. To make sure the compared traces cover similar spatial extents, we first clipped the longer line segment in a pair using the 500 m buffered MBR of the shorter line segment. In total, 85 glaciers have 159 same-day terminus traces across the period 2013–2022 (Fig. 9b). The average mean distance error for these glaciers is 32 ± 65 m, and 65 % of the analysed glaciers have a mean distance error between 10 and 30 m (Fig. 9c).

Since the mean distance error calculation only covers a limited number of glaciers over a short time period, we implemented an additional assessment by comparing the long-term calving front change rates of each glacier between these two data products. We used the same centreline approach, with the same centrelines, to generate the time series of the glacier calving front changes for the Moholdt et al. (2022) data product. Due to a mismatch in the marine-terminating glaciers included in these two different datasets, we analysed the common subset of 129 glaciers and compared their calving front change rates. Variations in observation densities over time among the glaciers in our dataset could introduce a potential bias in the linear regression analysis for estimating the long-term calving front change rates, which is not an issue for the Moholdt et al. (2022) data product with an annual temporal resolution. In order to facilitate the comparison of calving front change rates, we first converted the irregular calving front positions in our dataset to daily front change distances through linear interpolation, and then we calculated the monthly mean glacier calving front change distances. The calving front change rate was estimated by fitting a linear regression to the interpolated monthly front change time series. For each glacier, the calving front change rates were calculated within a common time window, which was defined by the overlapping time period between these two data products.

There is an excellent match between the spatial distribution of glacier calving front change rates obtained from the two products (Fig. 10a–b). The glacier calving front change rates derived from this study show a significant near-linear correlation with the glacier calving front change rates from Moholdt et al. (2022) ($R^2 = 0.98$, P-value <0.05) (Fig. 11a). The Morsnevbreen Glacier exhibits the highest advancing rate of around $-700\ \mathrm{m\ yr^{-1}}$ in both products (Fig. 11a). This glacier, known for its surging behaviour, experienced its most recent surging event between late 2016 and late 2018, during which it advanced approximately 5 km (Fig. A2a–c). At the Polakkbreen Glacier, the most significant calving front retreat rate is observed (Fig. 11a). During the period from 2016 to 2022, this glacier experienced a retreat of approximately 4 km (Fig. A2d–f). In addition, 92 % of the investigated glaciers show an absolute difference in calving front change rates of less than $25\ \mathrm{m\ yr^{-1}}$ between the two data products (Figs. 10c and 11b). The Storisstraumen Glacier in Austfonna Basin-3 exhibits the largest absolute difference in front change rate of $77\ \mathrm{m\ yr^{-1}}$ (Fig. A2g–i). Our data show a pronounced seasonal cycle in the calving front change of this glacier during the period 2014–2023 (black line in Fig. A2i). By contrast, the Moholdt et al. (2022) calving front measurements only record the most advanced location in September each year, resulting in an underestimation of the calving front advancing rate.

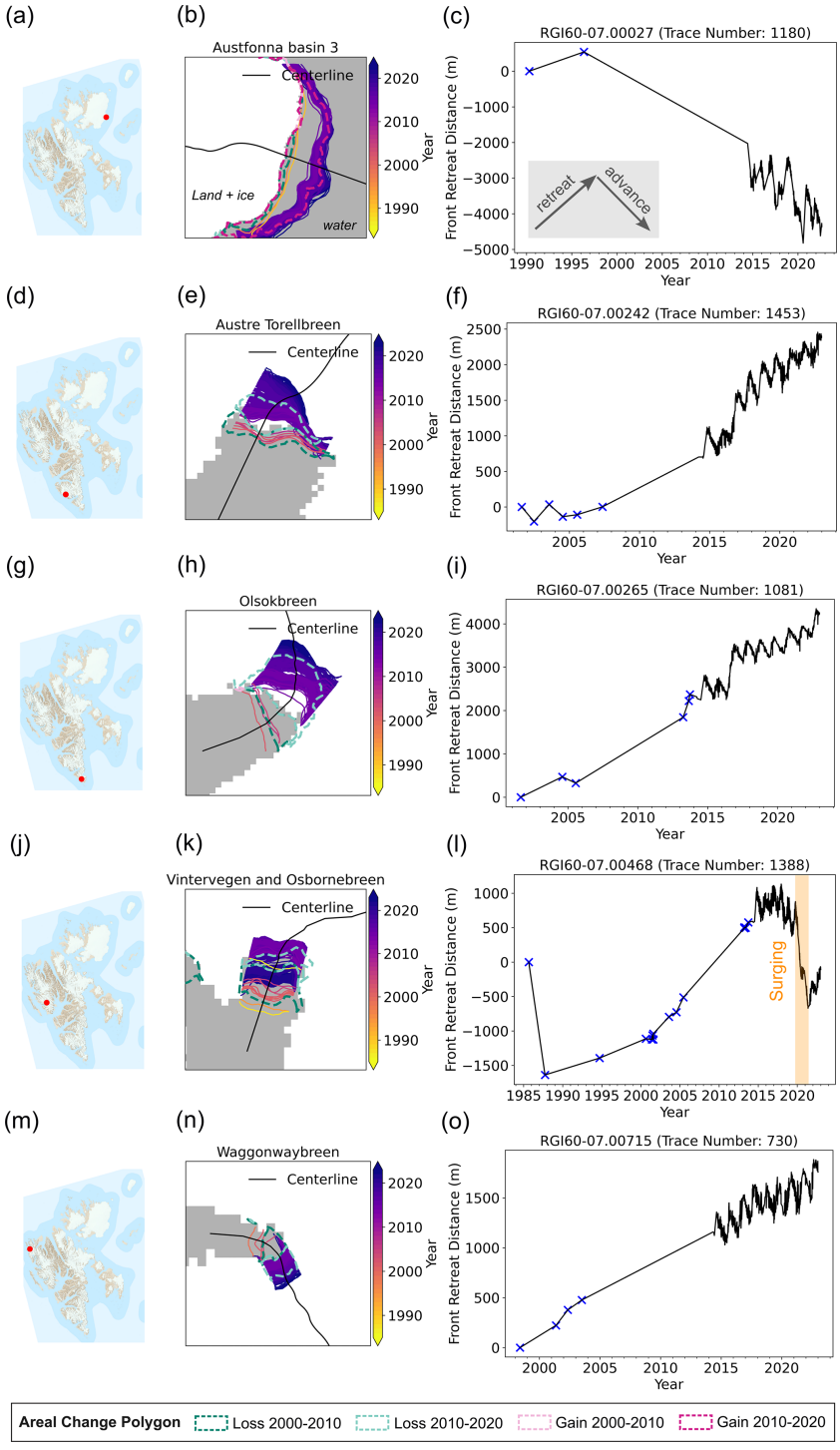


Figure 7. Examples of glacier calving front change time series of five different glaciers located across Svalbard. Red dots in panels **(a)**, **(d)**, **(g)**, **(j)**, and **(m)** show the locations of each glacier; the basemap is the S100 topographic raster data for Svalbard (https://data.npolar.no/dataset/44ca8c2a-22c2-49e8-a50b-972734f287e3, last access: 17 April 2023). In panels **(b)**, **(e)**, **(h)**, **(k)**, and **(n)**, coloured line segments are the glacier calving front traces mapped in this study for each glacier; they are overlaid with the 2000–2020 glacier areal change polygons (Kochtitzky and Copland, 2022) denoted by dashed coloured polygons (legend at the bottom of the figure), and the binary land-ice (white) and water mask (grey) generated in Sect. 2.3.2. Panels **(c)**, **(f)**, **(i)**, **(l)**, and **(o)** show the glacier calving front change time series in relation to the earliest calving front trace at each glacier (upward trend denotes retreating while downward trend denotes advancing as illustrated in **c**), blue crosses denote the calving front change observations before 2014. In panel **(l)**, the orange box denotes the glacier surging event that occurred around 2020.

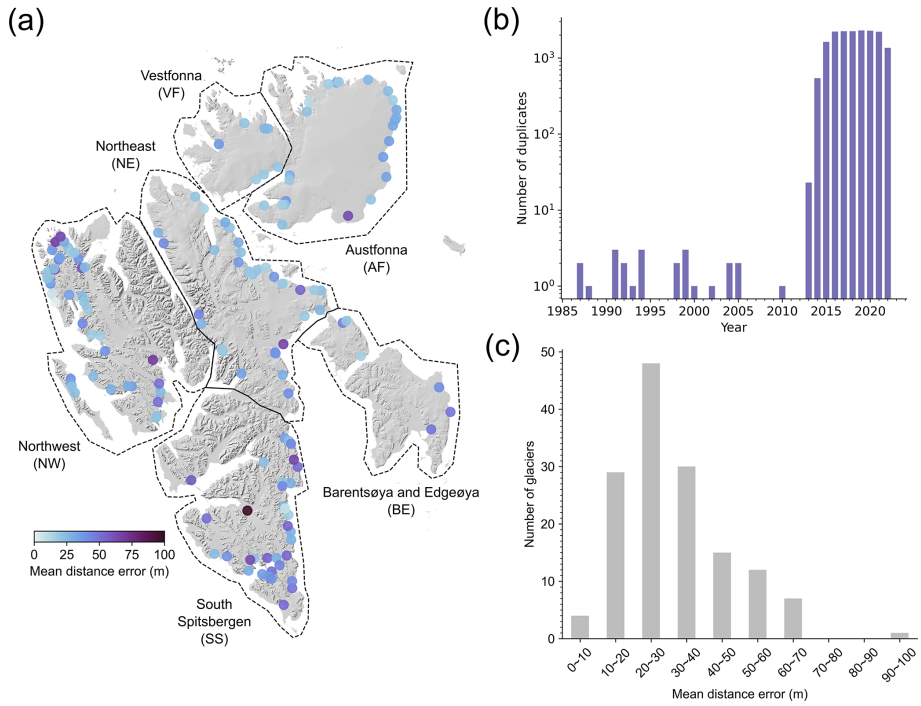


Figure 8. Calving front mapping mean distance error for 146 glaciers (3 glaciers do not have duplicated traces on the same day). **(a)** Spatial distribution of calving front mapping mean distance error of different marine-terminating glaciers. The background hillshade map is generated from the 50 m resolution Svalbard digital elevation model (DEM) (https://data.npolar.no/dataset/dce53a47-c726-4845-85c3-a65b46fe2fea, last access: 18 April 2023). **(b)** Temporal distribution of the same-day calving front trace duplicates. **(c)** Histogram of different mean distance error categories.

Table 3. Overview of two different calving front data products. "Type" indicates the type of calving front data provided in the data product. "Method" indicates how the dataset is produced. "No. glaciers" gives the number of presented glaciers. "No. mapped fronts" gives the total number of glacier calving front traces included in each data product.

Dataset	Data source	Type	Method	No. glaciers	No. mapped fronts	Time span	Temporal resolution
This study	Optical and SAR	Line	Neural network	149	124 919	1985–2023	Sub-weekly after 2014
Moholdt et al. (2022)	Optical	Line	Manually	202	2419	2008–2022	Annually

3.3 Spatial and temporal calving front variability in Svalbard

The spatial distribution of the different calving front change trends of the 149 marine-terminating glaciers included in the data product is shown in Fig. 12. The predominant trend among Svalbard's marine-terminating glaciers is retreat, where 123 glaciers (82.6 %) have been consistently retreating during the study period. Overall, 16 glaciers showed an advancing trend (not surging); most of these glaciers are located on the Vestfonna and Austfonna ice caps on the island of Nordaustlanet at the northeastern limit of the archipelago, where warm North Atlantic waters are less accessible (Fig. 12) (Skogseth et al., 2005). There are an additional 10 glaciers that displayed surge behaviour and they have a widespread distribution across different regions.

Svalbard is one of the most prominent regions of surge-type glaciers, with approximately 13 % showing this behaviour (Jiskoot et al., 2000). Using our extensive satellite data catalogue, we were able to capture the exact timing of surge-type events (Figs. 7j–l and 13) and identify surging events that are unknown from previous calving front data products (Kochtitzky and Copland, 2022; Moholdt et al., 2022). For example, Tunabreen is a quiescent-phase surge-type glacier which terminates in Temperfjorden, a shallow fjord with limited connection to the warm ocean currents (Luckman et al., 2015). During our study period, we observed two individual surging events at Tunabreen, one during 2002–2004, and the other during 2017–2019 (orange boxes in Fig. 13d). During both events, the Tunabreen calving front advanced more than 1.5 km in less than 2 years. By comparison, Moholdt et al. (2022) only identified the sec-

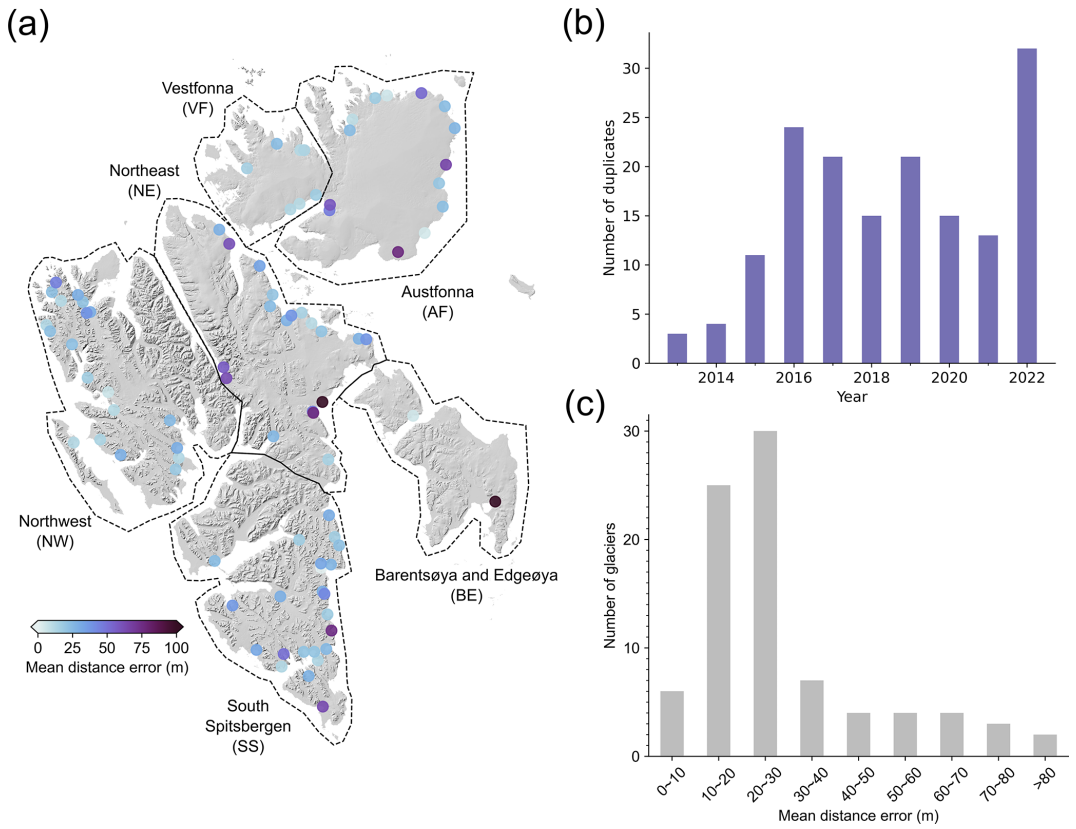


Figure 9. Calving front mapping mean distance error for 85 glaciers between data products generated in this study and by Moholdt et al. (2022). **(a)** Spatial distribution of calving front mapping mean distance error of different marine-terminating glaciers. The background hillshade map is generated from the 50 m resolution Svalbard digital elevation model (DEM) (https://data.npolar.no/dataset/dce53a47-c726-4845-85c3-a65b46fe2fea, last access: 18 April 2023). **(b)** Temporal distribution of the calving front traces mapped on the same day. **(c)** Histogram of different mean distance error categories.

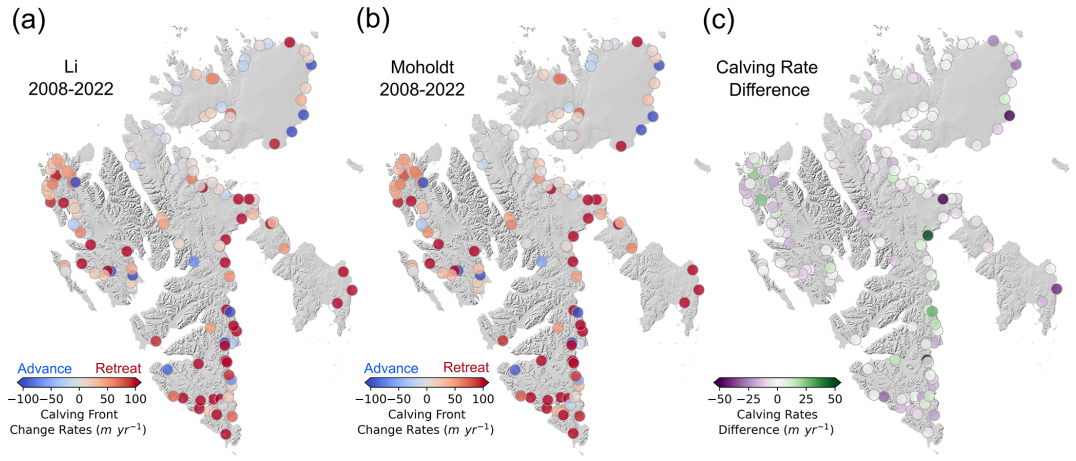


Figure 10. The calving front change rates between 2008 and 2022 for the calving front data product generated in this study **(a)**, the calving front data product by Moholdt et al. (2022) **(b)**, and the calving front change rate difference between these two calving front data products **(c)**. The background hillshade map is generated from the 50 m resolution Svalbard digital elevation model (DEM) (https://data.npolar.no/dataset/dce53a47-c726-4845-85c3-a65b46fe2fea, last access: 18 April 2023).

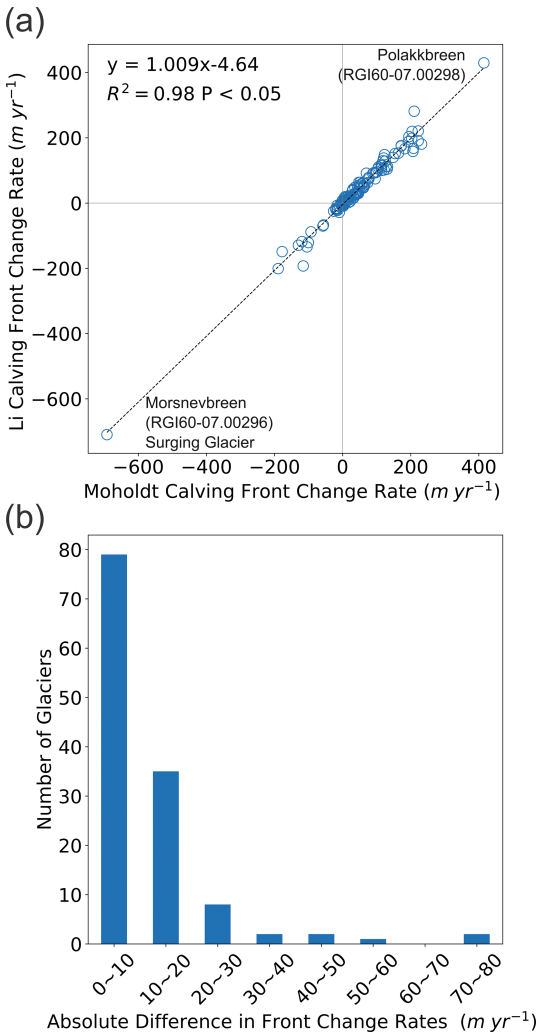


Figure 11. Comparison of glacier calving front change rates between product generated in this study and the Moholdt et al. (2022) calving front data product. Panel **(a)** shows the correlation between the glacier calving front change rates between these two different data products. Panel **(b)** shows the histogram of absolute difference in glacier front change rates between these two different calving front data products.

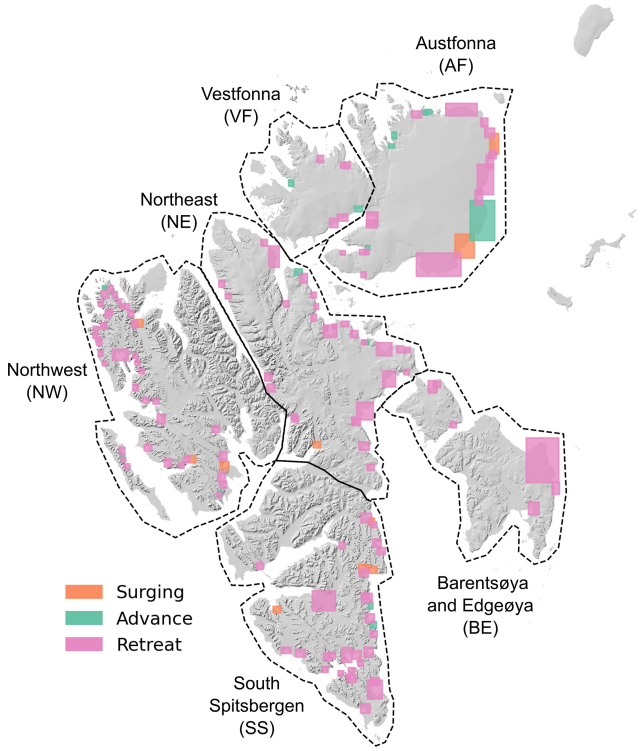


Figure 12. Spatial distribution of different calving front change trends of marine-terminating glaciers in Svalbard derived from the calving front data product generated in this study, and the main current circulation around the Svalbard archipelago (Skogseth et al., 2005; Misund et al., 2016). The orange, green, and pink polygons represent surging glaciers, non-surging-type advancing glaciers, and retreating glaciers, respectively. The background hillshade map is generated from the 50 m resolution Svalbard digital elevation model (DEM) (https://data.npolar.no/dataset/dce53a47-c726-4845-85c3-a65b46fe2fea, last access: 18 April 2023).

ond surging event (Fig. 13e), and they were also unable to capture the seasonal cycles of the calving events. This example demonstrates the power of our highly automated multi-sensor calving front mapping scheme, which can uncover previously unknown events in unprecedented detail and can aid future investigations on calving front dynamics and the mass balance of tidewater glaciers.

4 Discussion

Our calving front dataset of Svalbard marine-terminating glaciers during 1985–2023 is the first to provide calving front observations of large and comprehensive spatial coverage, high temporal resolution, and a long time span of 38 years. It not only captures the spatial pattern of evolving marine-terminating glacier calving fronts, but also provides insights at different time scales. This dataset can be used to study glacier mass balance, understand calving mechanisms, and predict glacier dynamics.

The calving front data product is mapped using the novel COBRA deep-learning model (Heidler et al., 2023). This model has been proven to outperform the previous calving front mapping models such as HED-UNet (Heidler et al., 2022), which was used for the IceLine Antarctic ice shelf front dataset (Baumhoer et al., 2023), CALFIN (Cheng et al., 2021), as well as the UNet model (Mohajerani et al., 2019). While the geomorphological features of tidewater glaciers in Svalbard and Greenland exhibit general similarities, it is important to note that the calving styles and neighbouring fjords can vary significantly among certain glaciers. Therefore, the CALFIN training dataset used in our model development

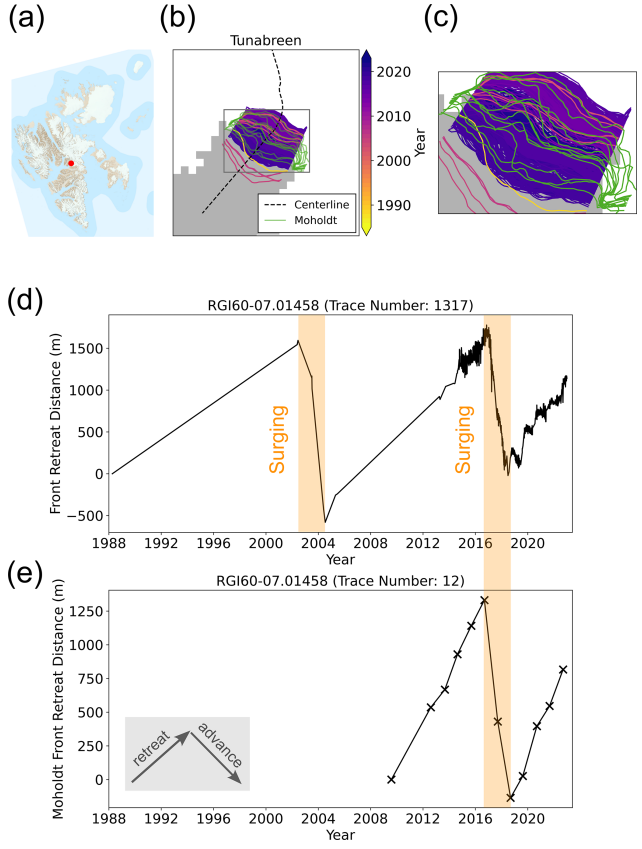


Figure 13. Calving front change time series of Tunabreen surging glacier (RGI60-07.01458). **(a)** Red dot shows the location of Tunabreen overlaid on the basemap from the S100 topographic raster data for Svalbard (https://data.npolar.no/dataset/44ca8c2a-22c2-49e8-a50b-972734f287e3, last access: 17 April 2023); **(b)** The coloured lines are the calving front traces derived in this study overlaid on the binary land-ice (white) and water mask (grey) generated in Sect. 2.3.2; the solid green lines are the calving front traces mapped in the Moholdt et al. (2022) data product; the glacier centreline is denoted by dashed black line. **(c)** The zoomed-in map of calving front traces inside the grey box in **(b)**. **(d)** The glacier calving front change time series included in this study, with the orange transparent boxes denoting two individual surging events. **(e)** The glacier calving front change time series from the Moholdt et al. (2022) data product; black crosses denote the calving front measurements.

may not be universally applicable to all Svalbard glaciers. To enhance the predictive capabilities of deep-learning models and simplify post-processing procedures, future research should focus on generating extensive training datasets for glacier calving fronts encompassing a wider range of geographical regions and glacier types.

Several external datasets were needed as inputs for the pre-processing and post-processing pipelines, including the Kochtitzky and Copland (2022) glacier front areal change polygon and the glacier centreline. The Kochtitzky and Copland (2022) areal change polygon serves the primary purpose of defining the glacier's bounding box for satellite image queries from GEE platform. Given that this areal change polygon only covers a limited period between 2000 and 2020, the fixed buffer length of 1.5 km used in Sect. 2.1 may not fully cover the entire calving front changes during 1985–2023. While this is less likely to be an issue in Svalbard given the relatively smaller scale and size of the marine-terminating glaciers, the buffer length will need to be adjusted when applying the processing pipeline to larger glaciers in different regions, such as the Greenland Ice Sheet. The glacier centreline is used in filtering out the abnormal front traces and producing the front change time series. Although only one centreline is used for each glacier, the centrelines are placed in areas with substantial calving front changes, making it effective and representative for filtering and quantifying the front changes over time.

The calving front changes of marine-terminating glaciers in our study are consistent with earlier observations by Kochtitzky and Copland (2022) and by Moholdt et al. (2022), although the temporal resolutions are different among these three products. The mean difference between our data product and the Moholdt et al. (2022) dataset is 32 ± 65 m, comparable to the calving front mapping uncertainty of 31 m in our dataset. In addition, the comparison of our glacier calving rates with the Moholdt et al. (2022) annual calving front data product shows an excellent match with $R^2 = 0.98$ during the period 2008–2022. The most significant mismatch in calving front change rate is located in Storisstraumen Glacier, and this is because the Moholdt et al. (2022) annual calving front dataset fails to capture the seasonal calving cycles. This example demonstrates the importance of considering seasonal calving front changes when estimating the long-term front change rates. Both datasets exhibit a clear and predominant trend of glacier retreat across Svalbard, in agreement with the Kochtitzky and Copland (2022) study of decadal glacier calving front change during 2000–2020, which shows that the net area change of glaciers in Svalbard is $-26.76 \pm 0.54\,\mathrm{km^2\,yr^{-1}}$. This spatial pattern was also reported by Geyman et al. (2022) by reconstructing DEMs using an archive of historical aerial imagery from 1936 and 1938. They showed that the mass balance in Svalbard during 1936–2010 was dominantly negative with an average thinning rate of $0.35 \pm 0.03\,\mathrm{m\,yr^{-1}}$. Glaciers in most of the regions experienced thinning rates exceeding $0.5\,\mathrm{m\,yr^{-1}}$, except the northeast Svalbard which remained stable during these 70 years.

Being able to assess calving front variability at multiple time scales is important in identifying drivers governing calving front changes and resolving mass balance estimations accurately (Benn and Åström, 2018; Rounce et al., 2023; Kochtitzky et al., 2022, 2023; Schuler et al., 2020; Luckman et al., 2015; Nuth et al., 2019; Strozzi et al., 2017; Cowton et al., 2018). Observations and theory show that increased calving can be driven by both atmospheric and oceanic warming. Increased surface melting and runoff can accelerate calving

through hydrofracturing of near-terminus crevasses. It can also increase subglacial discharge which, along with ocean warming, can drive submarine melting and accelerate terminus calving (Carr et al., 2013; Catania et al., 2020). Glacier calving processes in Svalbard, however, are not well understood due in part to a lack of comprehensive glacier calving front observations. Although Holmes et al. (2019) and Luckman et al. (2015) claimed that calving rates of marine-terminating glaciers in Svalbard vary strongly with ocean temperature, their results must be interpreted with caution – especially over large areas or long time scales – as they only used a small sample of glaciers ($n \leq 3$) within a short period of 2 years. The large number of investigated glaciers, along with the high temporal resolution and long time span (1985–2023) of our data product, provides a good basis for gaining new insights into the governing mechanisms in calving processes in Svalbard.

5 Code and data availability

The source code of COBRA model v1.0.0 and inference examples are accessible at https://github.com/khdlr/COBRA/releases/tag/v1.0.0 (last access: 10 January 2023), its DOI is https://doi.org/10.5281/zenodo.8407566 (Heidler, 2023). The Svalbard calving front dataset produced in this study is available at the Zenodo data repository: https://doi.org/10.5281/zenodo.10407266 (Li et al., 2023).

6 Conclusion

In this study, we produced a new high-resolution glacier calving front dataset, including 124 919 individual calving fronts, for 149 marine-terminating glaciers in Svalbard covering the period 1985–2023. This represents a significant increase in glacier calving front observation density compared to similar products. This data product was derived using automated processing methods developed in this study, which incorporate a novel deep-learning framework, multiple optical and SAR satellite images (Landsat, Terra-ASTER, Sentinel-1 and Sentinel-2) curated and downloaded via the Google Earth Engine platform, and a bespoke post-processing algorithm. The data product is validated with the latest Svalbard annual calving front dataset produced by Moholdt et al. (2022) by calculating the mean difference in calving front locations and comparing the calving front change rates over the same period of time. The results show a strong correlation in calving front change rates between the two products with an R^2 value of 0.98, while their mean difference is only 32 ± 65 m. In addition, our results show that calving front retreat has been dominant across most of Svalbard in the past four decades, except the northeast region comprising Vestfonna and Austfonna, consistent with the overall negative glacier mass balance identified in Svalbard. This new dataset will contribute to a better understanding of glacier calving front mechanisms and more accurate frontal ablation estimates in Svalbard. This is essential in calculating glacier mass balance and predicting the contribution to future sea-level rise, especially in the context of the ongoing Arctic warming.

Appendix A

Table A1. The Google Earth Engine (GEE) image collections for different satellites used in this study.

Satellite	GEE image collection
ASTER	ASTER/AST_L1T_003
Landsat-1	LANDSAT/LM01/C02/T1
Landsat-2	LANDSAT/LM02/C02/T1
Landsat-3	LANDSAT/LM03/C02/T1
Landsat-4	LANDSAT/LT04/C02/T1
Landsat-5	LANDSAT/LT05/C02/T1
Landsat-7	LANDSAT/LE07/C02/T1
Landsat-8	LANDSAT/LC08/C02/T1
Landsat-9	LANDSAT/LC09/C02/T1
Sentinel-2	COPERNICUS/S2_HARMONIZED
Sentinel-1	COPERNICUS/S1_GRD

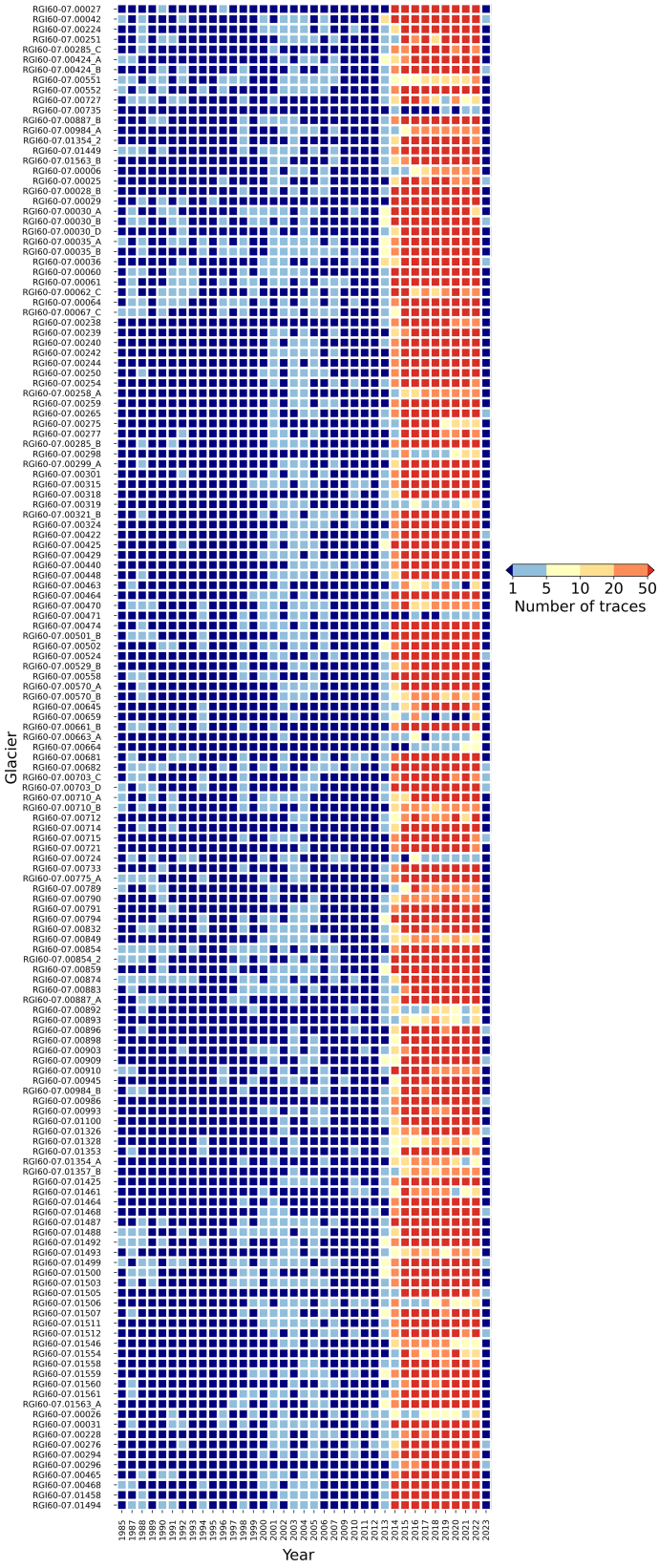


Figure A1. Heatmap of glacier traces of each marine-terminating glacier analysed in this study from 1985 to 2023 January. Each column represents one glacier, and each row represents 1 year ranging from 1985 to 2023. The colour corresponds to the number of traces for one glacier per year.

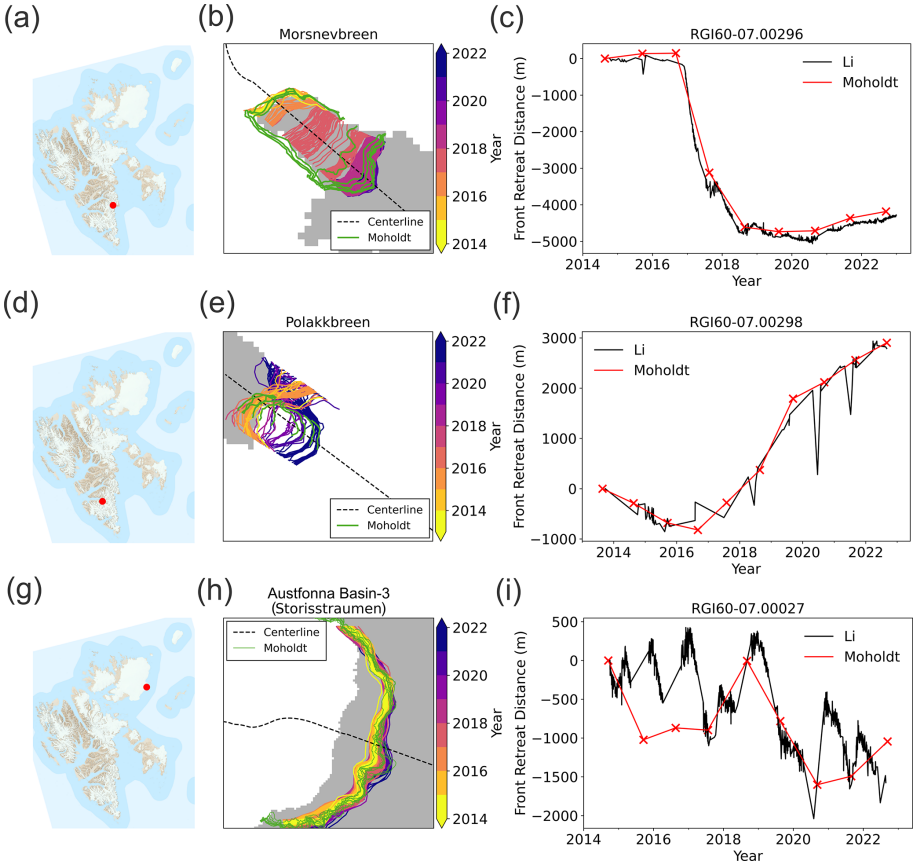


Figure A2. Examples of glacier calving front change comparison during a common time period between calving front data products generated in this study and by Moholdt et al. (2022) for Morsnevbreen Glacier (panels **a**–**c**), Polakkbreen Glacier (panels **d**–**f**), and Storisstraumen Glacier in Austfonna Basin-3 (panels **g**–**i**). Red dots in panels **(a)**, **(d)**, and **(g)** show the locations of each glacier, the basemap is the S100 topographic raster data for Svalbard (https://data.npolar.no/dataset/44ca8c2a-22c2-49e8-a50b-972734f287e3, last access: 17 April 2023). In panels **(b)**, **(e)**, and **(h)**, coloured line segments are the glacier calving front traces mapped in this study; they are overlaid with the calving front traces mapped in Moholdt et al. (2022) denoted by solid green lines, and the binary land-ice (white) and water mask (grey) generated in Sect. 2.3.2. Panels **(c)**, **(f)**, and **(i)** show the glacier calving front change time series in relation to the earliest calving front trace during the data comparison time window; solid black lines show the front change time series generated in this study and the solid red lines show the Moholdt et al. (2022) front change time series.

Author contributions. TL and JLB conceived the study. TL developed the data downloading, pre-processing and post-processing workflows, produced the results, and wrote the paper. KH developed the deep-learning model, trained the model, and contributed to data pre-processing. LM and AI contributed to the development of the post-processing pipeline. XXZ and JLB contributed to the interpretation of the results. All authors commented on the manuscript.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors.

Acknowledgements. This work is primarily funded by the European Union's Horizon 2020 research and innovation programme through the project Arctic PASSION (grant number: 101003472). Tian Li, Lichao Mou and Jonathan L. Bamber also received funding from the German Federal Ministry of Education and Research (BMBF) in the framework of the international future AI lab "AI4EO – Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond" (grant number: 01DD20001). Kon-

rad Heidler received funding from the German Federal Ministry for Economic Affairs and Climate Action in the framework of the "National Center of Excellence ML4Earth" (grant number: 50EE2201C). Xiao Xiang Zhu received funding from the Munich Center for Machine Learning (MCML). We would like to thank the editor Ken Mankoff and three anonymous reviewers for providing valuable comments that helped improve this study.

Financial support. This research has been supported by the Horizon 2020 (grant no. 101003472), the German Federal Ministry of Education and Research (grant no. 01DD20001), and the German Federal Ministry for Economic Affairs and Climate Action (grant no. 50EE2201C).

Review statement. This paper was edited by Ken Mankoff and reviewed by three anonymous referees.

References

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 Sens., 11, 2529, https://doi.org/10.3390/rs11212529, 2019.

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

Benn, D. I. and Åström, J. A.: Calving glaciers and ice shelves, Adv. Phys. X, 3, 1048–1076, https://doi.org/10.1080/23746149.2018.1513819, 2018.

Benn, D. I., Warren, C. R., and Mottram, R. H.: Calving processes and the dynamics of calving glaciers, Earth-Sci. Rev., 82, 143–179, https://doi.org/10.1016/J.EARSCIREV.2007.02.002, 2007.

Błaszczyk, M., Jacek, J., and Jon, O. H.: Tidewater Glaciers of Svalbard: Recent changes and estimates of calving fluxes, Polish Polar Res., 30, 85–142, https://opus.us.edu.pl/info/article/USL749ae2908280495bb99a7e046bb7cef1/ (last access: 9 August 2023), 2009.

Carr, J. R., Stokes, C. R., and Vieli, A.: Recent progress in understanding marine-terminating Arctic outlet glacier response to climatic and oceanic forcing: Twenty years of rapid change, Prog. Phys. Geogr., 37, 436–467, https://doi.org/10.1177/0309133313483163, 2013.

Carr, J. R., Stokes, C. R., and Vieli, A.: Threefold increase in marine-terminating outlet glacier retreat rates across the Atlantic Arctic: 1992–2010, Ann. Glaciol., 58, 72–91, https://doi.org/10.1017/AOG.2017.3, 2017.

Catania, G. A., Stearns, L. A., Moon, T. A., Enderlin, E. M., and Jackson, R. H.: Future Evolution of Greenland's Marine-Terminating Outlet Glaciers, J. Geophys. Res.-Earth, 125, e2018JF004873, https://doi.org/10.1029/2018JF004873, 2020.

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.

Chollet, F.: Xception: Deep Learning with Depthwise Separable Convolutions, in: Proc. – 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017, January 2017, 1800–1807, https://doi.org/10.1109/CVPR.2017.195, 2016.

Cook, A. J., Copland, L., Noël, B. P. Y., Stokes, C. R., Bentley, M. J., Sharp, M. J., Bingham, R. G., and van den Broeke, M. R.: Atmospheric forcing of rapid marine-terminating glacier retreat in the Canadian Arctic Archipelago, Sci. Adv., 5, eaau8507, https://doi.org/10.1126/SCIADV.AAU8507, 2019.

Cowton, T. R., Sole, A. J., Nienow, P. W., Slater, D. A., and Christoffersen, P.: Linear response of east Greenland's tidewater glaciers to ocean/atmosphere warming, P. Natl. Acad. Sci. USA, 115, 7907–7912, https://doi.org/10.1073/PNAS.1801769115, 2018.

Davies, T. M., Marshall, J. C., and Hazelton, M. L.: Tutorial on kernel estimation of continuous spatial and spatiotemporal relative risk, Stat. Med., 37, 1191–1221, https://doi.org/10.1002/sim.7577, 2018.

European Space Agency (ESA): Copernicus DEM – Global and European Digital Elevation Model (COP-DEM), ESA [data set], https://doi.org/10.5270/ESA-c5d3d65, 2021.

Geyman, E. C., van Pelt, W. J. J., Maloof, A. C., Aas, H. F., and Kohler, J.: Historical glacier change on Svalbard predicts doubling of mass loss by 2100, Nature, 601, 374–379, https://doi.org/10.1038/s41586-021-04314-4, 2022.

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 Syst. Sci. Data, 14, 4287–4313, https://doi.org/10.5194/essd-14-4287-2022, 2022.

Heidler, K.: khdlr/COBRA: v1.0.0, Zenodo [code], https://doi.org/10.5281/zenodo.8407566, 2023.

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 T. Geosci. Remote, 60, 4300514, https://doi.org/10.1109/TGRS.2021.3064606, 2022.

Heidler, K., Mou, L., Loebel, E., Scheinert, M., Lefèvre, S., and Zhu, X. X.: A Deep Active Contour Model for Delineating Glacier Calving Fronts, IEEE T. Geosci. Remote, 61, 5615912, https://doi.org/10.1109/TGRS.2023.3296539, 2023.

Holmes, F. A., Kirchner, N., Kuttenkeuler, J., Krützfeldt, J., and Noormets, R.: Relating ocean temperatures to frontal ablation rates at Svalbard tidewater glaciers: Insights from glacier proximal datasets, Sci. Rep., 9, 9442, https://doi.org/10.1038/s41598-019-45077-3, 2019.

Hugonnet, R., McNabb, R., Berthier, E., Menounos, B., Nuth, C., Girod, L., Farinotti, D., Huss, M., Dussaillant, I., Brun, F., and Kääb, A.: Accelerated global glacier mass loss in the early twenty-first century, Nature, 592, 726–731, https://doi.org/10.1038/s41586-021-03436-z, 2021.

Intergovernmental Panel on Climate Change: Climate Change 2021 – The Physical Science Basis, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, https://doi.org/10.1017/9781009157896, 2023.

Jiskoot, H., Murray, T., and Boyle, P.: Controls on the distribution of surge-type glaciers in Svalbard, J. Glaciol., 46, 412–422, https://doi.org/10.3189/172756500781833115, 2000.

Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, arXiv preprint, https://doi.org/10.48550/arXiv.1412.6980, 2014.

Kochtitzky, W. and Copland, L.: Retreat of Northern Hemisphere Marine-Terminating Glaciers, 2000–2020, Geophys. Res. Lett., 49, e2021GL096501, https://doi.org/10.1029/2021GL096501, 2022.

Kochtitzky, W., Copland, L., Van Wychen, W., Hugonnet, R., Hock, R., Dowdeswell, J. A., Benham, T., Strozzi, T., Glazovsky, A., Lavrentiev, I., Rounce, D. R., Millan, R., Cook, A., Dalton, A., Jiskoot, H., Cooley, J., Jania, J., and Navarro, F.: The unquantified mass loss of Northern Hemisphere marine-terminating glaciers from 2000–2020, Nat. Commun., 13, 5835, https://doi.org/10.1038/s41467-022-33231-x, 2022.

Kochtitzky, W., Copland, L., Van Wychen, W., Hock, R., Rounce, D. R., Jiskoot, H., Scambos, T. A., Morlighem, M., King, M., Cha, L., Gould, L., Merrill, P. M., Glazovsky, A., Hugonnet, R., Strozzi, T., Noël, B., Navarro, F., Millan, R., Dowdeswell, J. A., Cook, A., Dalton, A., Khan, S., and Jania, J.: Progress toward globally complete frontal ablation estimates of marine-terminating glaciers, Ann. Glaciol., 63, 143–152, https://doi.org/10.1017/aog.2023.35, 2023.

Li, T., Heidler, K., Mou, L., Ignéczi, Á., Zhu, X. X., and Bamber, J.: Calving Front Dataset for Marine-Terminating Glaciers in Svalbard 1985–2023, Zenodo [data set], https://doi.org/10.5281/zenodo.10407266, 2023.

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 T. Geosci. Remote, 60, 1–12, https://doi.org/10.1109/TGRS.2022.3208454, 2022.

Loebel, E., Scheinert, M., Horwath, M., Humbert, A., Sohn, J., Heidler, K., Liebezeit, C., and Zhu, X. X.: Calving front monitoring at sub-seasonal resolution: a deep learning application to Greenland glaciers, The Cryosphere Discuss. [preprint], https://doi.org/10.5194/tc-2023-52, in review, 2023.

Luckman, A., Benn, D. I., Cottier, F., Bevan, S., Nilsen, F., and Inall, M.: Calving rates at tidewater glaciers vary strongly with ocean temperature, Nat. Commun., 6, 8566, https://doi.org/10.1038/ncomms9566, 2015.

Maussion, F., Butenko, A., Champollion, N., Dusch, M., Eis, J., Fourteau, K., Gregor, P., Jarosch, A. H., Landmann, J., Oesterle, F., Recinos, B., Rothenpieler, T., Vlug, A., Wild, C. T., and Marzeion, B.: The Open Global Glacier Model (OGGM) v1.1, Geosci. Model Dev., 12, 909–931, https://doi.org/10.5194/gmd-12-909-2019, 2019.

McNabb, R. W. and Hock, R.: Alaska tidewater glacier terminus positions, 1948–2012, J. Geophys. Res.-Earth, 119, 153–167, https://doi.org/10.1002/2013JF002915, 2014.

Meredith, M., Sommerkorn, M., Cassotta, S., Derksen, C., Ekaykin, A., Hollowed, A., Kofinas, G., Mackintosh, A., Melbourne-Thomas, J., Muelbert, M. M. C., Ottersen, G., Pritchard, H., and Schuur, E. A. G.: Polar Regions, IPCC Special Report on the Ocean and Cryosphere in a Changing Climate, Cambridge University Press, 203–320, https://doi.org/10.1017/9781009157964.005, 2019.

Misund, O. A., Heggland, K., Skogseth, R., Falck, E., Gjøsæter, H., Sundet, J., Watne, J., and Lønne, O. J.: Norwegian fisheries in the Svalbard zone since 1980. Regulations, profitability and warming waters affect landings, Polar Sci., 10, 312–322, https://doi.org/10.1016/J.POLAR.2016.02.001, 2016.

Mohajerani, Y., Wood, M., Velicogna, I., and Rignot, E.: Detection of glacier calving margins with convolutional neural networks: A case study, Remote Sens., 11, 74, https://doi.org/10.3390/rs11010074, 2019.

Moholdt, G., Maton, J., Majerska, M., and Kohler, J.: Annual frontlines of marine-terminating glaciers on Svalbard, https://data.npolar.no/dataset/d60a919a-9cc8-4048-9686-df81bfdc2338 (last access: 16 May 2023), 2022.

Müller, M.: Dynamic Time Warping, Inf. Retr. Music Motion, 69–84, https://doi.org/10.1007/978-3-540-74048-3_4, 2007.

Murray, T., Scharrer, K., Selmes, N., Booth, A. D., James, T. D., Bevan, S. L., Bradley, J., Cook, S., Llana, L. C., Drocourt, Y., Dyke, L., Goldsack, A., Hughes, A. L., Luckman, A. J., and McGovern, J.: Extensive Retreat of Greenland Tidewater Glaciers, 2000–2010, Arctic, Antarct. Alp. Res., 47, 427–447, https://doi.org/10.1657/AAAR0014-049, 2015.

Noël, B., Jakobs, C. L., van Pelt, W. J. J., Lhermitte, S., Wouters, B., Kohler, J., Hagen, J. O., Luks, B., Reijmer, C. H., van de Berg, W. J., and van den Broeke, M. R.: Low elevation of Svalbard glaciers drives high mass loss variability, Nat. Commun., 11, 4597, https://doi.org/10.1038/s41467-020-18356-1, 2020.

Nordli, Ø., Wyszyński, P., Gjelten, H. M., Isaksen, K., Łupikasza, E., Niedźwiedź, T., and Przybylak, R.: Revisiting the extended Svalbard Airport monthly temperature series, and the compiled corresponding daily series 1898–2018, Polar Res., 39, https://doi.org/10.33265/POLAR.V39.3614, 2020.

Nuth, C., Moholdt, G., Kohler, J., Hagen, J. O., and Kääb, A.: Svalbard glacier elevation changes and contribution to sea level rise, J. Geophys. Res., 115, F01008, https://doi.org/10.1029/2008JF001223, 2010.

Nuth, C., Kohler, J., König, M., von Deschwanden, A., Hagen, J. O., Kääb, A., Moholdt, G., and Pettersson, R.: Decadal changes from a multi-temporal glacier inventory of Svalbard, The Cryosphere, 7, 1603–1621, https://doi.org/10.5194/tc-7-1603-2013, 2013.

Nuth, C., Gilbert, A., Köhler, A., McNabb, R., Schellenberger, T., Sevestre, H., Weidle, C., Girod, L., Luckman, A., and Kääb, A.: Dynamic vulnerability revealed in the collapse of an Arctic tidewater glacier, Sci. Rep., 9, 5541, https://doi.org/10.1038/s41598-019-41117-0, 2019.

Oppenheimer, M., Glavovic, B. C., Hinkel, J., van de Wal, R. S. W., Magnan, A. K., Abd-Elgawad, A., Cai, R., Cifuentes-Jara, M., DeConto, R. M., Ghosh, T., Hay, J., Isla, F., Marzeion, B., Meyssignac, B., and Sebesvari, Z.: Sea Level Rise and Implications for Low-Lying Islands, Coasts and Communities, in: IPCC Special Report on the Ocean and Cryosphere in a Changing Climate, edited by: Pörtner, H.-O., Roberts, D. C., Masson-Delmotte, V., Zhai, P., Tignor, M., Poloczanska, E., Barrett, K., Seneviratne, S. I., and Macbean, N., Cambridge University Press, Cambridge, UK and New York, NY, USA, 321–445, 2019.

Pfeffer, W. T., Arendt, A. A., Bliss, A., Bolch, T., Cogley, J. G., Gardner, A. S., Hagen, J. O., Hock, R., Kaser, G., Kienholz, C., Miles, E. S., Moholdt, G., Mölg, N., Paul, F., Radić, V., Rastner, P., Raup, B. H., Rich, J., Sharp, M. J., Andreassen, L. M., Bajracharya, S., Barrand, N. E., Beedle, M. J., Berthier, E., Bhambri, R., Brown, I., Burgess, D. O., Burgess, E. W., Cawkwell, F., Chinn, T., Copland, L., Cullen, N. J., Davies, B., De Angelis, H., Fountain, A. G., Frey, H., Giffen, B. A., Glasser, N. F., Gurney, S. D., Hagg, W., Hall, D. K., Haritashya, U. K., Hartmann, G., Herreid, S., Howat, I., Jiskoot, H., Khromova, T. E., Klein, A.,

Kohler, J., König, M., Kriegel, D., Kutuzov, S., Lavrentiev, I., Le Bris, R., Li, X., Manley, W. F., Mayer, C., Menounos, B., Mercer, A., Mool, P., Negrete, A., Nosenko, G., Nuth, C., Osmonov, A., Pettersson, R., Racoviteanu, A., Ranzi, R., Sarikaya, M. A., Schneider, C., Sigurdsson, O., Sirguey, P., Stokes, C. R., Wheate, R., Wolken, G. J., Wu, L. Z., and Wyatt, F. R.: The Randolph Glacier Inventory: a globally complete inventory of glaciers, J. Glaciol., 60, 537–552, https://doi.org/10.3189/2014JOG13J176, 2014.

Rantanen, M., Karpechko, A. Y., Lipponen, A., Nordling, K., Hyvärinen, O., Ruosteenoja, K., Vihma, T., and Laaksonen, A.: The Arctic has warmed nearly four times faster than the globe since 1979, Commun. Earth Environ., 3, 168, https://doi.org/10.1038/s43247-022-00498-3, 2022.

RGI Consortium: Randolph Glacier Inventory (RGI) – A Dataset of Global Glacier Outlines: Version 6.0, NSIDC [data set], https://doi.org/10.7265/N5-RGI-60, 2017.

Rounce, D. R., Hock, R., Maussion, F., Hugonnet, R., Kochtitzky, W., Huss, M., Berthier, E., Brinkerhoff, D., Compagno, L., Copland, L., Farinotti, D., Menounos, B., and McNabb, R. W.: Global glacier change in the 21st century: Every increase in temperature matters, Science, 379, 78–83, https://doi.org/10.1126/science.abo1324, 2023.

Schuler, T. V., Kohler, J., Elagina, N., Hagen, J. O. M., Hodson, A. J., Jania, J. A., Kääb, A. M., Luks, B., Małecki, J., Moholdt, G., Pohjola, V. A., Sobota, I., and Van Pelt, W. J. J.: Reconciling Svalbard Glacier Mass Balance, Front. Earth Sci., 8, 156, https://doi.org/10.3389/feart.2020.00156, 2020.

Serreze, M. C. and Barry, R. G.: Processes and impacts of Arctic amplification: A research synthesis, Glob. Planet. Change, 77, 85–96, https://doi.org/10.1016/J.GLOPLACHA.2011.03.004, 2011.

Skogseth, R., Haugan, P. M., and Jakobsson, M.: Watermass transformations in Storfjorden, Cont. Shelf Res., 25, 667–695, https://doi.org/10.1016/J.CSR.2004.10.005, 2005.

Strozzi, T., Kääb, A., and Schellenberger, T.: Frontal destabilization of Stonebreen, Edgeøya, Svalbard, The Cryosphere, 11, 553–566, https://doi.org/10.5194/tc-11-553-2017, 2017.

van Pelt, W. J. J., Pohjola, V. A., Pettersson, R., Ehwald, L. E., Reijmer, C. H., Boot, W., and Jakobs, C. L.: Dynamic Response of a High Arctic Glacier to Melt and Runoff Variations, Geophys. Res. Lett., 45, 4917–4926, https://doi.org/10.1029/2018GL077252, 2018.

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., Catania, G., and Trugman, D. T.: AutoTerm: an automated pipeline for glacier terminus extraction using machine learning and a "big data" repository of Greenland glacier termini, The Cryosphere, 17, 3485–3503, https://doi.org/10.5194/tc-17-3485-2023, 2023.

A.4 Developing and Testing a Deep Learning Approach for Mapping Retrogressive Thaw Slumps

Reference

I. Nitze, K. Heidler, S. Barth, and G. Grosse, "Developing and testing a deep learning approach for mapping retrogressive thaw slumps," *Remote Sensing*, vol. 13, no. 21, p. 4294, 2021. DOI: 10.3390/rs13214294

Copyright

remote sensing

MDPI

Article

Developing and Testing a Deep Learning Approach for Mapping Retrogressive Thaw Slumps

Ingmar Nitze [1,*], Konrad Heidler [2,3], Sophia Barth [1,4] and Guido Grosse [1,4]

1 Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, 14473 Potsdam, Germany; sophia.barth@awi.de (S.B.); guido.grosse@awi.de (G.G.)
2 Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), 82234 Weßling, Germany; konrad.heidler@dlr.de
3 Data Science in Earth Observation, Technical University of Munich (TUM), 80333 Munich, Germany
4 Institute of Geosciences, University of Potsdam, 14476 Potsdam, Germany
* Correspondence: ingmar.nitze@awi.de

Abstract: In a warming Arctic, permafrost-related disturbances, such as retrogressive thaw slumps (RTS), are becoming more abundant and dynamic, with serious implications for permafrost stability and bio-geochemical cycles on local to regional scales. Despite recent advances in the field of earth observation, many of these have remained undetected as RTS are highly dynamic, small, and scattered across the remote permafrost region. Here, we assessed the potential strengths and limitations of using deep learning for the automatic segmentation of RTS using PlanetScope satellite imagery, ArcticDEM and auxiliary datasets. We analyzed the transferability and potential for pan-Arctic upscaling and regional cross-validation, with independent training and validation regions, in six different thaw slump-affected regions in Canada and Russia. We further tested state-of-the-art model architectures (UNet, UNet++, DeepLabv3) and encoder networks to find optimal model configurations for potential upscaling to continental scales. The best deep learning models achieved mixed results from good to very good agreement in four of the six regions (maxIoU: 0.39 to 0.58; Lena River, Horton Delta, Herschel Island, Kolguev Island), while they failed in two regions (Banks Island, Tuktoyaktuk). Of the tested architectures, UNet++ performed the best. The large variance in regional performance highlights the requirement for a sufficient quantity, quality and spatial variability in the training data used for segmenting RTS across diverse permafrost landscapes, in varying environmental conditions. With our highly automated and configurable workflow, we see great potential for the transfer to active RTS clusters (e.g., Peel Plateau) and upscaling to much larger regions.

Keywords: deep learning; image segmentation; permafrost thaw; semantic segmentation; disturbances; computer vision; automation; PlanetScope; thermo-erosion; ArcticDEM; landslides

Citation: Nitze, I.; Heidler, K.; Barth, S.; Grosse, G. Developing and Testing a Deep Learning Approach for Mapping Retrogressive Thaw Slumps. *Remote Sens.* **2021**, *13*, 4294. https://doi.org/10.3390/rs13214294

Academic Editor: Michael Lim

Received: 22 September 2021
Accepted: 11 October 2021
Published: 26 October 2021

1. Introduction

The changing climate of the Arctic, with both measured and projected air temperatures and precipitation rapidly increasing [1,2], has a significant impact on permafrost [3–5]. As permafrost soils store about twice the amount of carbon as that found in the atmosphere [6,7], permafrost thaw and resulting carbon feedbacks are expected to have a significant impact on the global climate [8]. Rising permafrost ground temperatures have been observed across almost the entire Arctic permafrost region [3]. As a result of warming, permafrost becomes more vulnerable to disturbances of [9] and degradation in ground ice-rich landscapes due to thermokarst and thermo-erosion.

Retrogressive thaw slumps (RTS) are typical landforms related to processes of rapidly thawing and degrading hillslope permafrost [10]. Although these mass-wasting processes have been observed in different Arctic regions in the past decades [11–13], many recent

field and remote sensing studies found increasing occurrence and faster progression in various permafrost regions [12,14–18].

As RTS typically have a small size (<10 ha, with a few exceptions reaching up to ~1 km^2), as well as a wide range of appearances and dynamics, their detection and monitoring on the regional to continental scale would require globally available imagery at sufficiently high spatial and temporal resolutions. Their formation is bound to specific environmental and permafrost conditions, such as ice-rich permafrost and sloped terrains [10,12,19], thus limiting their presence to regional clusters. Particularly, regions with massive amounts of buried ice, as preserved in the moraines of former glaciations [17,20,21], or regions with thick syngenetic ice-wedges in yedoma permafrost [22,23], or with very fine grained marine deposits that were raised above sea level following deglaciation and thus formed very icy epigenetic permafrost, can be prone to RTS development [18]. Furthermore, increasing temperatures and precipitation have likely caused the increased formation and growth of RTS [21,24].

Fairly well-studied regions for the occurrence of thaw slumps are typically clustered and located in former ice-marginal regions of the Laurentide Ice Sheet in NW Canada, most notably the Peel Plateau [17,21] and Banks Island [16], or moraines of formerly glaciated mountain ranges, e.g., the Brooks Range in northern Alaska [20,25]. Intensively studied regions in Siberia include the Yamal Peninsula [13,26], Kolguev Island [27], Bolshoy Lyakhovsky Island [22] and the Yana Basin with its famous Batagaika mega slump [14,28]. However, the latter is, atypically, not part of a larger cluster of RTS. The total quantity and distribution of RTS in the Arctic remains unknown.

Several remote sensing studies have used very high-resolution (VHR) satellite data, but RTS are typically delineated manually, which is a labor-intensive task and therefore prohibitive for larger regions. The use of airborne [29,30] or UAV data [31] to survey small areas with RTS is becoming more popular. These datasets allow for the creation of elevation data and multiple observations, thus providing a basis for more automated approaches [29–31]. Highly automated approaches, which will be required to map RTS across larger regions and multiple time steps, are fairly scarce so far. Nitze et al. [32] used a random forest machine learning approach to map RTS and other permafrost disturbances, such as lake dynamics and wildfire, on Landsat data across four large north–south transects in the Arctic covering ~2.2 million km^2. For the indirect detection of RTS and thaw-related erosion features, Lara et al. [33] measured changes in lake color as a proxy for rapid thermo-erosion dynamics in a watershed-scale study in NW Alaska using Landsat. However, the coarse resolution of Landsat (30 m) proved to be a highly limiting factor in detecting RTS features accurately [32]. A combination of Landsat and Sentinel-2 imagery was used to assess RTS dynamics with the LandTrendr disturbance detection algorithm over a ~8 million km^2 region of East Siberia for a 20-year period from 2000 to 2019 [34].

Automated approaches applied to higher-resolution data (better than 5 m ground resolution), such as high-resolution RapidEye and PlanetScope imagery or very high-resolution DigitalGlobe/Maxar imagery, pose specific challenges for image classification and specifically object detection. On such data, pixel-based approaches are no more feasible, and object-based image approaches (OBIA) need to be applied [35]. Traditionally, this has been accomplished with the segmentation followed by classification of image objects. Over the past few years, deep learning (DL) techniques have grown in popularity for object detection or segmentation in imagery of any kind, e.g., bio-medical images or everyday photography.

In remote sensing, DL approaches are also growing in popularity [36] for typical applications such as image segmentation and classification, due to their ability to take spatial context into account. This includes, e.g., the mapping of landslides [37–39]. Furthermore, DL-based image segmentation has been particularly applied on VHR data, such as Worldview, GeoEye, etc., to automatically detect comparably small objects, such as buildings [40–42] or individual trees [43]. Due to many DL algorithms, such as Mask R-CNN,

requiring a fixed amount of input bands, e.g., one or three, and to avoid overfitting, several studies have focused on input band selection and optimization [44–46].

In permafrost remote sensing, deep learning applications are very scarce so far. They have been applied for mapping and segmenting ice-wedge polygons [47–49] and for detecting infrastructure across the Arctic permafrost region [50]. DL for detecting and tracking RTS was used by Huang et al. [51,52], who tested the applicability of the DeepLabv3+ DL architecture for detecting and monitoring RTS on the Tibetan Plateau using Planet data. They received a high detection quality similar to manual digitization [52], which enabled them to track RTS in space and time within a confined region.

Based on these promising achievements, we here aim to:

(1) test the feasibility of applying DL methods on PlanetScope and auxiliary data to detect and map RTS across different Arctic permafrost regions;

(2) identify the particular advantages and challenges;

(3) discuss the further requirements for using AI-based techniques to eventually map RTS across the circum-Arctic permafrost zone.

2. Materials and Methods

2.1. Study Regions

We selected six different sites across the Arctic in Canada and Russia that are affected by RTS (Figure 1; Table 1). These locations were chosen to contain a sufficient number of RTS, and to represent a broad variety of environmental conditions (sparse tundra to taiga) and geographic settings (RTS at coast, river, or lake shores, hillslopes, and moraines). Study sites with a spatially extensive occurrence of RTS (e.g., Horton Delta, Banks Island, Kolguev Island) were each split into two subsets. All sites/subsets have an area of 100 km^2 (10 × 10 km) to ensure the best possible comparison and normalization to each other.

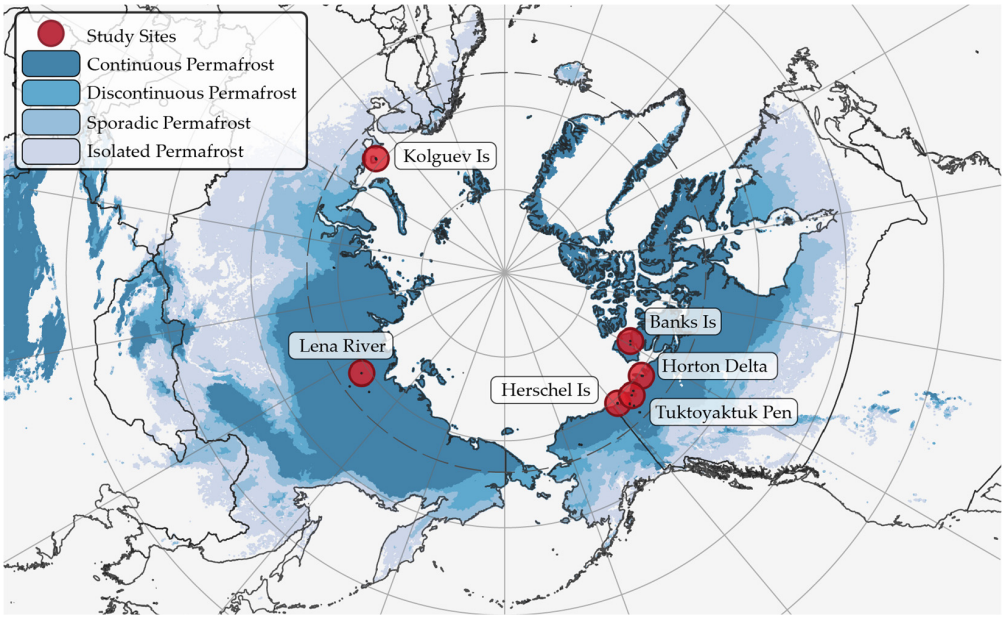


Figure 1. Overview map of study sites and permafrost extent based on (Obu et al., 2018).

Table 1. Study sites with center coordinates, region, and number of used Planet images.

Study Site	Center Coordinates	Region	# of Images	# of Image Dates
Banks Island 01	119.50° W; 72.84° N;	NW Canada	12	5
Banks Island 02	118.20° W; 73.04° N	NW Canada	15	4
Herschel Island	139.00° W; 69.60° N	NW Canada	10	5
Horton Delta 01	126.75° W; 69.75° N;	NW Canada	10	4
Horton Delta 02	126.60° W; 69.64° N	NW Canada	13	6
Kolguev Island 01	48.35° E; 69.22° N	NW Siberia	29	14
Kolguev Island 02	48.51° E; 69.35° N	NW Siberia	20	8
Lena River	124.40° E; 69.12° N	E Siberia	47	22
Tuktoyaktuk Pen.	133.80° W; 69.12° N	NW Canada	19	9

2.1.1. Banks Island

The Banks Island study site consists of two subsets and is located in the eastern RTS-rich part of Banks Island in NW Canada (see Figures A1 and A2). This region is characterized by glacial moraine deposits (Jesse Moraine) of the former Laurentide Ice Sheet, which contains buried massive ground ice [16,53]. The region is subject to massive permafrost degradation as indicated by strong ice-wedge degradation [54] and abundant RTS [16], which mostly form along lake shores and valley slopes. The vegetation is sparse tundra according to the Circum-Arctic Vegetation Map (CAVM) subzone C [55]. The selected site has some of the largest and most active RTS known globally (see Figure A1d). The region has rolling terrain with an abundance of lakes and river valleys. Modeled ground temperatures are −14 to −15 °C [56].

2.1.2. Herschel Island

This study site covers large parts of Herschel Island in NW Canada (see Figure A3). The Herschel Island site contains large highly active RTS, which have been frequently studied over the past decade [12,57]. Similar to many other RTS-rich sites in NW Canada, Herschel Island is located along the margins of the Laurentide Ice Sheet. The substrate is dominated by permafrost with massive buried glacial ice remnants [58]. The vegetation is dominated by shrubby tundra (erect dwarf shrub tundra) of CAVM Zone E [55]. Due to the rolling hilly nature of the island, there are many small stream catchments, but only a few smaller lakes and ponds. Thaw slumps are predominantly located on the SE shore. Modeled ground temperatures are −5 to −6 °C [56].

2.1.3. Horton Delta

This study site consists of two subsets and is located just south of the Horton River delta in NW Canada at a steep cliff on the Beaufort Sea coast (see Figures A4 and A5). This region was located at the front of a Laurentide Ice Sheet lobe and is known to be affected by RTS [21]. The site is characterized by rolling terrain with steep coastal cliffs and partially deeply incised valleys. Vegetation here is classified as dwarf shrub tundra of CAVM subzone D/E [55]. Lakes are very sparse, but larger valleys with rivers are present within this site. Thaw slumps are predominantly located on top of the coastal cliffs. Smaller RTS are also found along steep valley slopes in close proximity to the coast. Modeled ground temperatures are −7 to −8 °C [56].

2.1.4. Kolguev Island

Kolguev is an island off the coast of Arctic European Russia. It is characterized by ice-rich permafrost with tabular ice [27]. Vegetation here is dominated by Tundra of CAVM Zone D [55]. The study site has a rolling terrain with steep coastal bluffs. RTS are most abundant on these coastal bluffs in the NW of the island. Lakes are very sparse in this region (see Figures A6 and A7). Modeled ground temperatures are 0 to 1 °C [56], though the presence of RTS and therefore ground-ice suggests lower temperatures.

2.1.5. Lena River

This study site is located in the lower reaches of the Lena River on the east side of the river close to the foothills west of the Verkhoyansk Mountain Range in northeastern Siberia. This site is likely a terminal moraine of an ancient outlet glacier from the mountain range, which underwent several glaciations during the Quaternary period [59]. The glacial history of this region is not documented in detail. Vegetation here is boreal forest, and the region is lake-rich. RTS typically formed along the lake's shores. Former stabilized RTS are also abundant and mostly covered by dense shrubs (see Figure A8). The presence of RTS in this region is only sparsely documented in the literature [32]. Modeled ground temperatures are −7 to −8 °C [56].

2.1.6. Tuktoyaktuk Peninsula

This region is located on the Tuktoyaktuk Peninsula in NW Canada. It is a rolling, glacially (Laurentide Ice Sheet) shaped lowland with massive ground ice [19,60]. The vegetation here is shrubby tundra of CAVM Zone E [55] close to the tundra–taiga ecotone. This region has a large abundance of lakes [61,62]. Thaw slumps typically form along lake shores (see Figure A9). Modeled ground temperatures are −6 to −7 °C [56].

2.2. Data

For training data collection, as well as model training, validation and inference, we used a variety of data. Our primary data source was the PlanetScope [63] multispectral optical data for the years 2018 and 2019. We further used additional datasets, such as the ArcticDEM [64] and Tasseled Cap Landsat Trends [32]. Furthermore, for collecting ground truth, we additionally used the ESRI and Google Satellite layers.

2.2.1. PlanetScope

We used PlanetScope satellite images [63] as our primary data source. PlanetScope data are acquired by a constellation of more than 120 satellites in orbit. They have a spatial resolution of 3.15 m and four spectral bands in the visual (red, green, blue; RGB) and near-infrared (NIR) wavelengths. The high number of satellites in orbit allows for sub-daily temporal resolution, particularly at high-latitudes, where data overlap becomes increasingly dense for satellites following a polar orbit [65]. However, non-obstructed views of the ground are severely limited, particularly in high-latitude coastal regions, due to persistent cloud cover and cloud shadows, haze, and long snow periods. Furthermore, the generally low sun elevations in high-latitude environments can lead to low signal-to-noise ratios.

For data selection, we applied the following selection criteria: maximum 10% cloud cover, 90% area coverage, and an observation period from 1 June until 30 September during the years 2018 and 2019. Furthermore, we selected image dates by visual inspection to ensure consistent temporal sampling, where possible. As cloud-free periods (the main limiting factor) tended to be temporally clustered, we omitted several clear sky image dates within short periods (e.g., five consecutive days with clear skies), as these will not provide additional value for training the model. The image dates and IDs are indicated in Figure 2 and Supplementary Table S1. Due to further satellite launches, the number of PlanetScope images increased significantly over our observation period. Thus, available imagery was rather sparse before 2019, but became increasingly abundant after that.

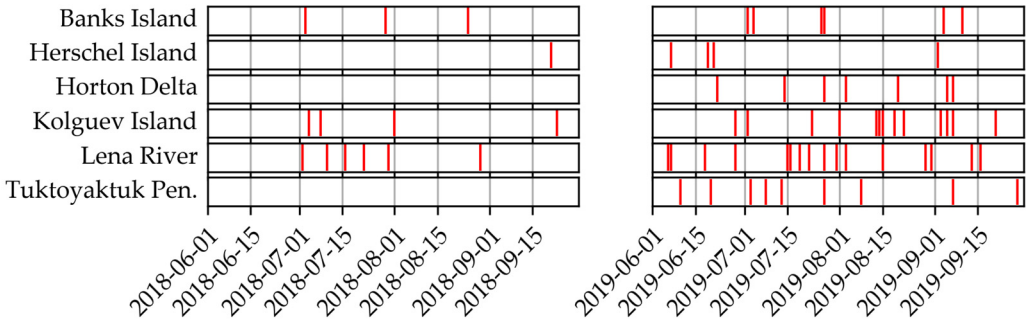


Figure 2. Temporal distribution of input data by study region.

Finally, we downloaded data through the *porder* download program [66] and Planet Explorer interface. We chose the "analytic_sr,udm2" bundle, which includes surface reflectance data, udm (unusable data mask), udm2 and metadata files. We chose to clip output data automatically to the respective AOI extents, which allowed us to optimize the allocated data quota and to ensure the completeness of all ground truth datasets. Finally, we calculated the Normalized Difference Vegetation Index (NDVI) for each scene as an additional input layer. We used the udm2 data mask to mask out remaining clouds, shadows and snow/ice in the PlanetScope and all auxiliary datasets.

2.2.2. Arctic DEM

We used the ArcticDEM [64] (version 3, Google Earth Engine Dataset: "UMN/PGC/ArcticDEM/V3/2m_mosaic") to calculate slope and detrended elevation data. The ArcticDEM is available for all land areas north of 60° latitude, but contains minor data gaps. We calculated the relative (detrended) elevation by subtracting the mean elevation within a circular window (structuring element) with a diameter of 50 pixels (100 m). The relative elevation was used to determine the local pixel position within the surroundings and to remove the influence of the regional elevation. Finally, we rescaled the relative elevation values with an offset of 50 and factor of 300 to minimize the size of data of the unsigned Integer16 type. Furthermore, we calculated the slope in degrees. For both calculations we used the *ee.Terrain.slope* function in the Google Earth Engine (GEE).

We downloaded the data (relative elevation and slope) for the training sites (buffered by 5 km) from GEE with the native projection (NSIDC Sea Ice Polar Stereographic North, EPSG: 3413) and a spatial resolution of 2 m. We chose GEE over the original data portal due to the simpler accessibility of data, as well as its capacity for slope calculation and process automation. After downloading, all individual tiles were merged into virtual mosaics using *gdalbuildvrt* to simplify data handling and permit efficient data storage. We later reprojected both datasets, elevation and slope, to the projection, spatial resolution and image extent of individual PlanetScope scenes using *gdalwarp* within our automated processing pipeline (see Figure 3).

2.2.3. TCVIS

To introduce a decadal-scale multi-temporal dataset into the analysis, we used the temporal trend datasets of Tasseled Cap indices of Landsat data (Collection 1, Tier 1, Surface Reflectance), based on previous work [67,68]. For the period from 2000 to 2019, we filtered Landsat data to scenes with a cloud cover of less than 70% and masked clouds, shadows and snow/ice based on available masking data.

We calculated the Tasseled Cap indices [69], brightness (TCB), greenness (TCG), and wetness (TCW) for each individual scene. Then we calculated the linear trend for each index over time, scaled to 10 years. Finally, we truncated the slope values of all three indices to a range of −0.12 to 0.12 and transformed the data to an unsigned integer data range (0 to 255) to minimize storage use. The resulting data were stored as a publicly readable GEE ImageCollection asset ("users/ingmarnitze/TCTrend_SR_2000-2019_TCVIS").

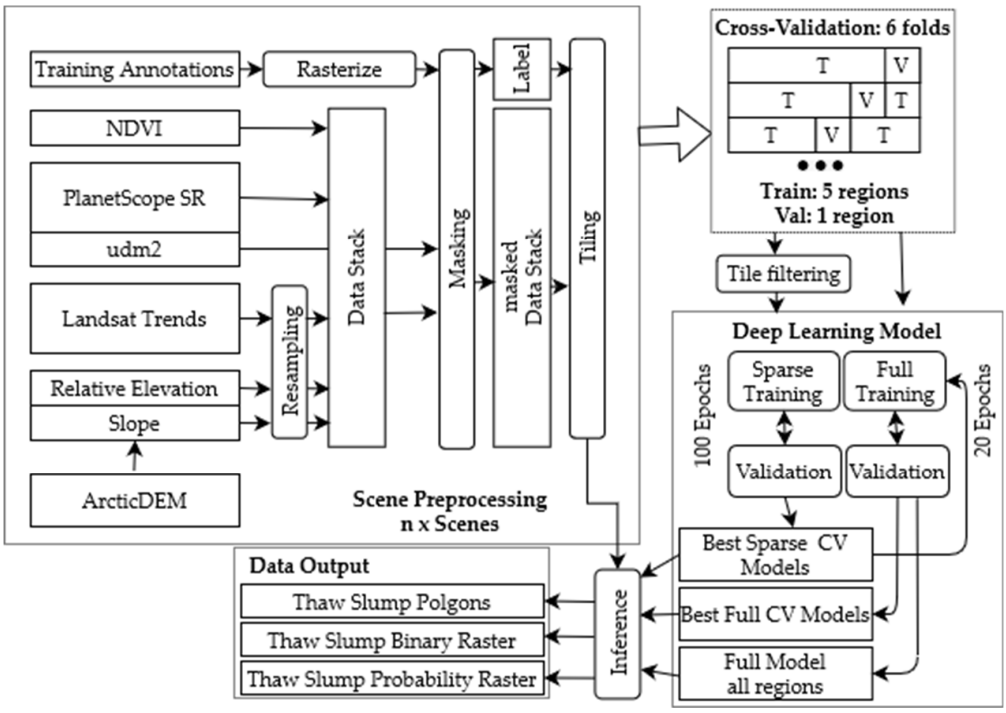


Figure 3. Flowchart of the full processing pipeline.

2.3. Methods

2.3.1. Slump Digitization

We created ground truth datasets for training and validation by manual digitization in QGIS 3.10 [70]. We used the individual PlanetScope scenes (see Section 2.2.1) as the primary data source. We digitized each available image individually. Accordingly, we have multi-temporal information of RTS in all study regions. The same physical RTS may have different polygon shapes on different dates due to the physical change of the RTS (e.g., growth), presence of snow, its location on the edge of the imagery, geolocation inaccuracies, or slightly inconsistent digitization (see below).

Furthermore, we used auxiliary data to better understand landscape morphology and landscape dynamics, when interpreting potential RTS features. These auxiliary data are the ArcticDEM and multitemporal TCVIS (Landsat Tasseled Cap Trend) data, streamed through the Google Earth Engine Plugin (https://github.com/gee-community/qgis-earthengine-plugin, v0.0.2) in QGIS. Furthermore, additional VHR imagery publicly available in ESRI and Google satellite base layers was accessed and used for mapping through the QuickMapServices Plugin in QGIS [71]. The VHR imagery was used solely for guidance in order to better identify the ground objects at a higher resolution than the 3 m PlanetScope imagery.

Labeling went through two iterations to ensure the highest data quality. In the first step, a trained person manually digitized potential thaw slumps that matched selected criteria. During this iteration, unclear cases were discussed with a second trained person. The criteria for manually outlining RTS in the data were:

1. little or no vegetation, surrounded by vegetation;
2. presence of a headwall;
3. "blue" signature of TCVIS layer, a transition from vegetation to wet soil;
4. visible depression in ArcticDEM and derived slope dataset;
5. visible thaw slump disturbance in VHR imagery;
6. snow was considered as not being part of the RTS.

In the next step the second person checked each individual thaw slump object and confirmed, edited, removed or added new polygons. In this procedure, we closely follow the RTS digitizing guidelines set out by Segal et al. [72].

Although the datasets went through several iterations, oftentimes it was challenging to determine whether the slumps were still active or already stabilized, and therefore whether they needed to be included in or excluded from the process. Furthermore, while actively eroding upper parts of thaw slumps were easy to delineate due to the presence of a headwall, the lower scar zone and debris tongues were typically more challenging to delineate due to unclear boundaries. Overall we digitized 2172 thaw slump polygons. Please find more details in Table 2. The digitized polygons are made freely accessible (see Data Availability Statement).

Table 2. Study sites with total number of detected RTS and number of individual RTS per date.

Study Site	# of Total Individual RTS Objects	# of Individual RTS per Date [1, 2]	Median Object Size (m^2)
Banks Island 01	397	65–103	22,032
Banks Island 02	151	24–53	22,203
Herschel Island	148	15–40	5175
Horton Delta 01	180	36–52	5562
Horton Delta 02	354	35–67	7981
Kolguev Island 01	44	3–5	78,786
Kolguev Island 02	275	25–41	13,840
Lena River	238	5–13	14,470
Tuktoyaktuk Pen.	385	30–55	2229

[1] Total image size may be different between dates, e.g., incomplete coverage. [2] PlanetScope data have some image overlap, which may lead to (partially) duplicated vectors.

2.3.2. Deep Learning Model

General Setup

For the data preprocessing, model training, validation, and inference we developed a highly automated processing pipeline to ensure the highest possible level of automation, reproducibility and transferability (see Figure 3). It is easily configurable with configuration files, which allow us to define the key processing parameters, such as dataset (train, val, test), data sources (see Table 3), DL model architecture and encoder, model depth, and many more. Our processing chain is based on the pytorch deep learning framework [73] within the python programming language. Furthermore, we relied on several additional packages for specific tasks (see below).

Table 3. List of model input data layers, with preprocessing status, native resolution and number of bands.

Input Data	Raw/Derived	Native Resolution (m)	# Bands
PlanetScope Scene (SR)	Raw	3	3
PlanetScope NDVI	Derived	3	1
ArcticDEM relative elevation	Derived	2	1
ArcticDEM slope	Derived	2	1
TCVIS	Derived	30	3

The processing was split into three main steps: first, data preprocessing; second, model training and validation; third, model inference.

The code is tracked and documented in a git repository (see code). We used version 0.4.1 for the training and validation. We performed the inference on version 0.5.2, which included bug fixes related to inference, compared to version 0.4.1.

Hardware

We ran our model training and inference on virtual machines equipped with a shared and virtualized NVIDIA GV100GL GPU (Tesla V100 PCIe 32 GB). The VM was allocated with 16 GB GPU RAM, 8x Intel(R) Xeon(R) Gold 6230 CPU, 128 GB RAM and fast storage.

Augmentation

In order to increase its size and to introduce more variety into the training dataset, the input imagery was augmented in several ways. Since satellite imagery is largely independent of orientation, the images were randomly mirrored along their horizontal and vertical axes, as well as being rotated by multiples of 90°. Randomly blurring some input images during training further improved model robustness. Augmentation increased the training set size eight-fold. Image augmentation was conducted and implemented using the *Albumentations* python library [74]. Each augmentation type was randomly applied with a probability of 50% per image.

Model Architecture

For the pixel-wise classification of images, semantic segmentation models offer an efficient approach to combining local information and contextual clues. For our model architecture we evaluated some network architectures commonly used for semantic segmentation. These segmentation architectures consist of an encoder network and a decoder. Successful image classification architectures are commonly used as encoders, as these can efficiently extract general image features. Therefore, we evaluated three ResNet [75] architectures (Resnet34, Resnet50 and Resnet101) as possible encoders for our network. Decoders are currently undergoing the most innovation in semantic segmentation, and thus vary a lot from architecture to architecture. Here, we evaluated three approaches, namely, UNet [76], UNet++ [77] and DeepLabv3 [78]. The model architectures are based on the implementation of the *segmentation_models_pytorch* package (https://github.com/qubvel/segmentation_models.pytorch, v0.2.0).

Training Details and Hyperparameters

All trained models were initialized randomly. For optimizing the training loss, the Adam optimizer was used, setting the hyperparameters as suggested by Kingma and Ba [79], namely $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\varepsilon = 10{-}8$. We used a learning rate of 0.001 and batch size of 256 $\times$ 256 pixels with a 25 pixel overlap. The stack height was set to 6. We used Focal Loss as the loss function after testing different options.

Cross-Validation: Data Setup

We performed a thorough regional cross-validation (CV), where we used 5 regions for training and the 1 remaining region for validation. We rotated through all regions so that each region was used as the validation set once, which totals six folds. Regions with multiple subsets (Banks Island, Horton, Kolguev) were treated as one for validation. For each regional fold we performed a parameter grid search over each of the three model architectures and three encoders. Each model has nine input layers in total (see Table 3). The complete dataset consists of 11863 image tiles, of which 1317 contain RTS.

For computing the classification and segmentation performance, we used the following pixel-wise metrics: overall accuracy and Cohen's kappa for the overall classification performance. Furthermore, we used the class-specific metrics Intersection over Union (IoU), precision, recall, and F1 for only the positive class (RTS) to determine the class-specific performance and balance. We calculated all metrics for each individual epoch for the training and validation set, which provided information about the model's gradual performance improvement. Validation was automatically carried out during the model training phase. Training and validation metrics for each epoch are automatically stored in the output logs. Model performance evaluation was carried out in this configuration.

Furthermore, for the final model evaluation and inter-comparison, we also sorted each run by performance from best to worst.

We carried out the CV training and validation scheme in two steps. First, for each of the 54 configurations we ran the training for 100 epochs on sparse training sets. To train the model we only used tiles with targets (RTS), thus undersampling the background/non-RTS class in order to (1) reduce class imbalance and (2) speed up the training process. Finally, we added a second training stage of 20 epochs for the best calculated model (highest IoU score) for the three best regions with the full training set, including a high proportion of negative/non-slump tiles. Non-slump tiles are all image tiles that do not include any RTS, and which comprise the vast majority of tiles, due to the sparse occurrence of RTS. This second step was carried out to place further emphasis on training negative samples, as the initial tests showed a strong overestimation of slumps in stable regions.

Inference for Spatial Evaluation

We carried out inference runs to determine the spatial patterns and segmentation capabilities of the trained models. For this purpose, we applied three different strategies.

(1a) We used the best model (highest IoU score) of each cross-validation training scheme and ran the inference for the validation sites. This strategy provided us with completely unseen/independent information on the spatial transferability with strengths and weaknesses of the models.

(1b) We used the fully trained model (sparse and full training) of the best configuration per region and carried out the inference for each region.

(2) We used the fully trained model (all regions) on the best overall configuration, and ran inference on all the input images and PlanetScope imagery of the study regions from 2020 and 2021. This recent imagery was not clipped to the 10 × 10 km study site size. Thaw slumps outside the study site boundaries were therefore unknown to the trained models, and could serve as independent objects from a different period, yet within the proximity of the trained region.

For all inference runs, we chose a standard configuration of 1024 × 1024 pixels tile-size with an overlap of 128 pixels. For merging the tile overlap we selected a soft-margin approach, wherein the overlapping areas of adjacent tiles are blended linearly.

The model creates three different output layers (Figure 3, Table 4). First, a probability (p-value) raster layer (GTiff), which contains the probability of each pixel belonging to the RTS class. Second, a binary raster mask (GTiff) with a value of 1 for RTS locations (p-value > 0.5). Third, a polygon vector file (ESRI Shapefile) with predicted RTS, converted from the binary raster mask.

Table 4. List of model inference output data layers.

Output Data	Format	Resolution (m)
Polygon vector	ESRI Shapefile	-
Binary raster	GTiff	3
Probability raster	GTiff	3

3. Results

3.1. AI Model Performance

3.1.1. Train/Test/Cross-Validation Performance

The applied AI segmentation models performed similarly, but with certain differences and slightly diverging performances. In all configurations, the training performance increased with increasing epochs (Figures 9a and A10). Furthermore, the validation performance exceeded training metrics from the beginning, and typically plateaued from around 50 epochs. The good early validation performance compared to the training shows the effect of augmentation and indicates low overfitting.

3.1.2. Regional Comparison

The regionally stratified cross-validation on the sparse training sets highlighted the regional differences in thaw slumps across the Arctic with regard to environmental conditions, data quality and data availability. Overall, regional differences were more pronounced than model specifics or configurations, such as architecture and encoder. In the following, named regions indicate the validation (unseen) dataset, while the remaining regions were used for training (regionally stratified cross-validation).

The Lena validation set achieved the best results (best model, see Table 5) with maximum IoU scores of 0.58 (UNet++ Resnet34), followed by Horton (0.55, UNet++ Resnet101), Kolguev (0.48, UNet++ Resnet101) and Herschel (0.38, DeepLabv3 Resnet34). Banks Island (UNet++ Resnet50) achieved a maximum IoU of 0.39, but this deteriorated quickly, seemingly due to strong overfitting. Tuktoyaktuk (UNet++ Resnet101) only achieved a maximum IoU of 0.15, with very little improvement even after several epochs (Figure 4a).

Table 5. Regional performance of best sparse models. U++: UNet++; DLv3: DeepLabv3; Rn34: Resnet34; Rn50: Resnet50; Rn101: Resnet101. IoU1/Prec1/Recall1/$F_1$1: Metrics of best sparse regional CV model. IoU5: 5th best model of 100. IoU10: 10th best model of 100.

Study Site	Model Config.	IoU1	IoU5	IoU10	Prec1	Recall1	$F_1$1
Banks Island	U++Rn50	0.39	0.13	0.08	0.80	0.38	0.52
Herschel	DLv3Rn34	0.39	0.33	0.32	0.50	0.63	0.56
Horton	U++Rn101	0.55	0.54	0.51	0.78	0.77	0.71
Kolguev	U++Rn101	0.48	0.45	0.43	0.67	0.63	0.64
Lena	U++Rn34	0.58	0.51	0.50	0.83	0.65	0.73
Tuktoyaktuk	U++Rn50	0.15	0.09	0.08	0.42	0.18	0.25

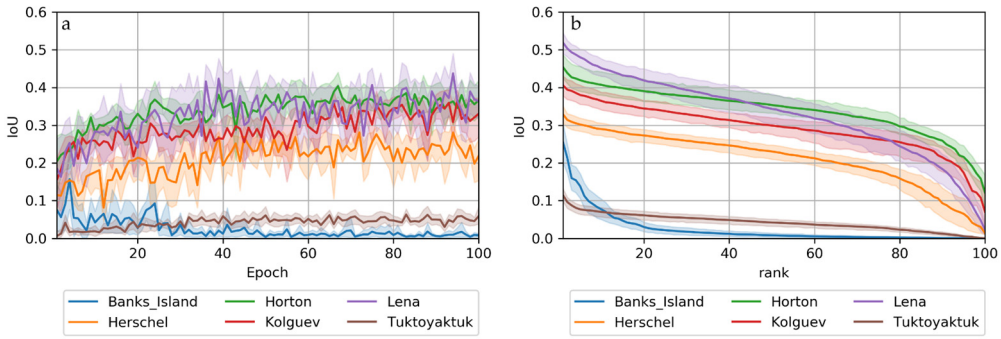


Figure 4. Mean and standard deviation of IoU scores per site sorted by (**a**) epoch and (**b**) scores (best to worst). n = 9 (3 sites × 3 encoders) per region.

Although the best models per region performed similarly, the mean/ensemble performance of all models per region typically differed much more significantly (Figure 4). For many regions, individual models behaved differently in terms of volatility and learning success.

The maximum accuracies/scores of validation sets typically plateaued after around 40 epochs with almost all configurations (Figure 4a), except for Banks Island. Banks Island achieved individual IoU scores > 0.2 during early epochs, and these converged quickly towards zero during later epochs, which suggests insufficient spatial transferability likely due to overfitting. Tuktoyaktuk suffered from low scores throughout the entire training period, with only little variation in its IoU of around <0.1. The difference in segmentation performance between the best and next models was typically small, except for Banks Island, as shown in the sorted model performance illustrations (Figure 4b).

3.1.3. Model Configurations

Among the tested configurations, including architectures and encoders, we only observed little differences in segmentation performance. However, overall, UNet++ outper-

formed UNet and DeepLabv3 consistently in this particular area (Figure 5). The choice of encoder only produced minor differences, but overall, simpler models (Resnet 34 > Resnet50 > Resnet101) resulted in slightly better IoU scores than more complex encoders (Resnet34: meanIoU = 0.33; Resnet50: meanIoU = 0.32; Resnet101: meanIoU = 0.31). In some individual cases, complex encoders (Resnet101) outperformed simpler encoders (e.g., Horton or Kolguev) (see Figure A10).

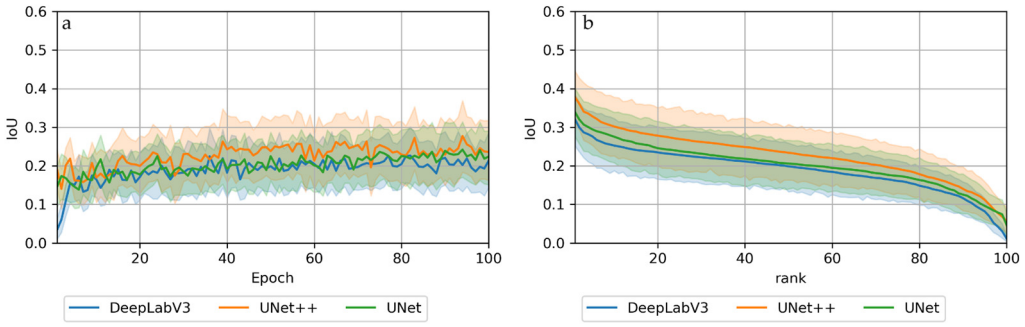


Figure 5. Mean and standard deviation of IoU scores per architecture sorted by (**a**) epoch and (**b**) scores (best to worst). n = 18 (6 sites × 3 encoders) per architecture.

3.1.4. Computation Performance

In all configurations, UNet was the fastest model with the least hardware requirements. UNet++ was ~60% slower (factor 1.6) than UNet, while DeepLabv3 improved training times by a factor of ~2.3 compared to UNet. The hardware requirements for GPU memory were in line with those for processing times, with UNet requiring the least resources, followed by UNet++ and DeepLabv3.

3.2. Inference/Spatial Model Output

Regional Cross-Validation

(1a) Sparse models: The sparse trained models, using only image tiles with positive samples (RTS), produced results ranging from unsatisfactory to acceptable (see Figure 4), depending on region and model used. Figures 6–8 (left column) show that the detection of thaw slumps produces mixed results, with strong variation depending on the input image. Decision boundaries are often fuzzy, with probability values (p-values) between 20 and 80% of being an RTS, as predicted by the model. Many non-slump areas, e.g., flat uplands or water bodies, were classified as thaw slumps in numerous instances, thus creating an abundance of false-positives in these settings.

(1b) Fully trained models: After adding further training epochs with the entire dataset, using predominantly negative samples, the results were visually improved, with more distinct decision boundaries. This manifests in the improved precision but reduced recall (see Figure 9). However, the full accuracy metrics IoU and F1 increased (sparse/full; Horton Delta: IoU:0.62/ 0.55, F1: 0.76/0.71), stayed the same (Lena River: IoU:0.65/0.66, F1: 0.73/0.74) or even decreased (Kolguev Island: IoU:0.48/0.38, F1: 0.64/0.55) depending on the specific site.

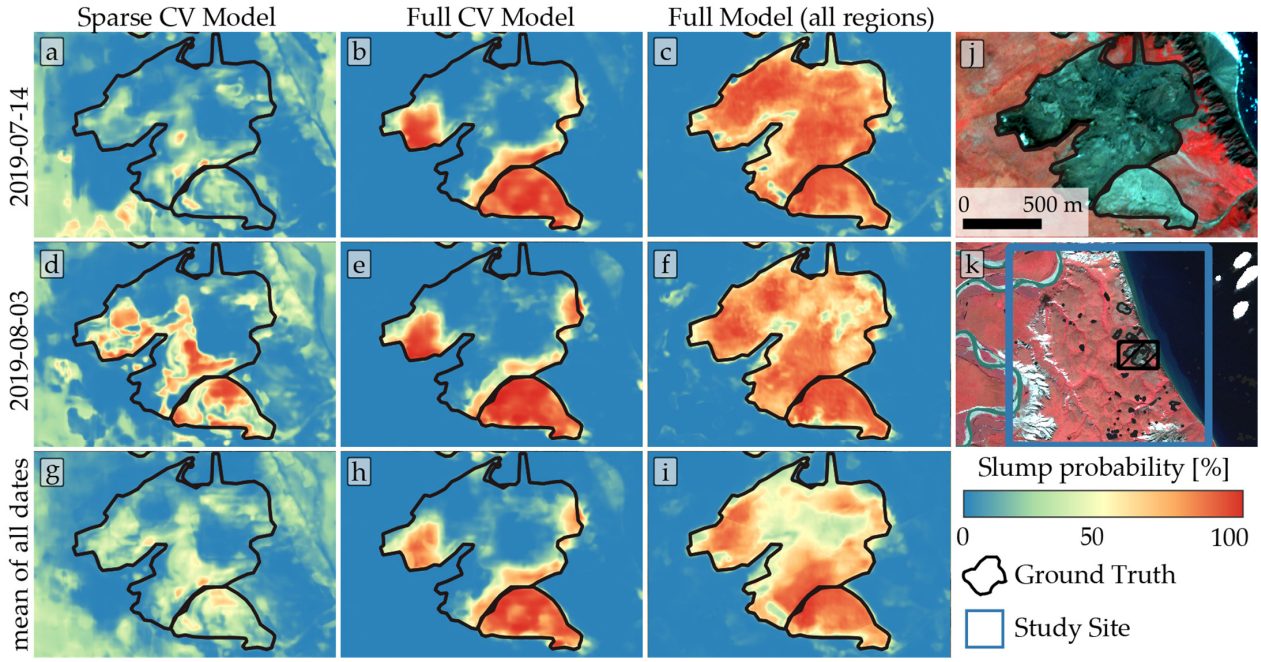


Figure 6. Detection results in subset of Horton Delta 01 study site with the modeled RTS probability on two different image dates ((**a**–**c**): 14 July 2019; (**d**–**f**): 03 August 2019) and the mean of all dates (**g**–**i**) as well as three different models ((**a**,**d**,**g**): sparse cross-validation model; (**b**,**e**,**h**): full cross-validation model; (**c**,**f**,**i**): full model trained on all regions). The subset of a multispectral false-color PlanetScope image (NIR-R-G) with ground truth is shown in panel (**j**). Panel (**k**) shows the approximate location of the subset within the study region.

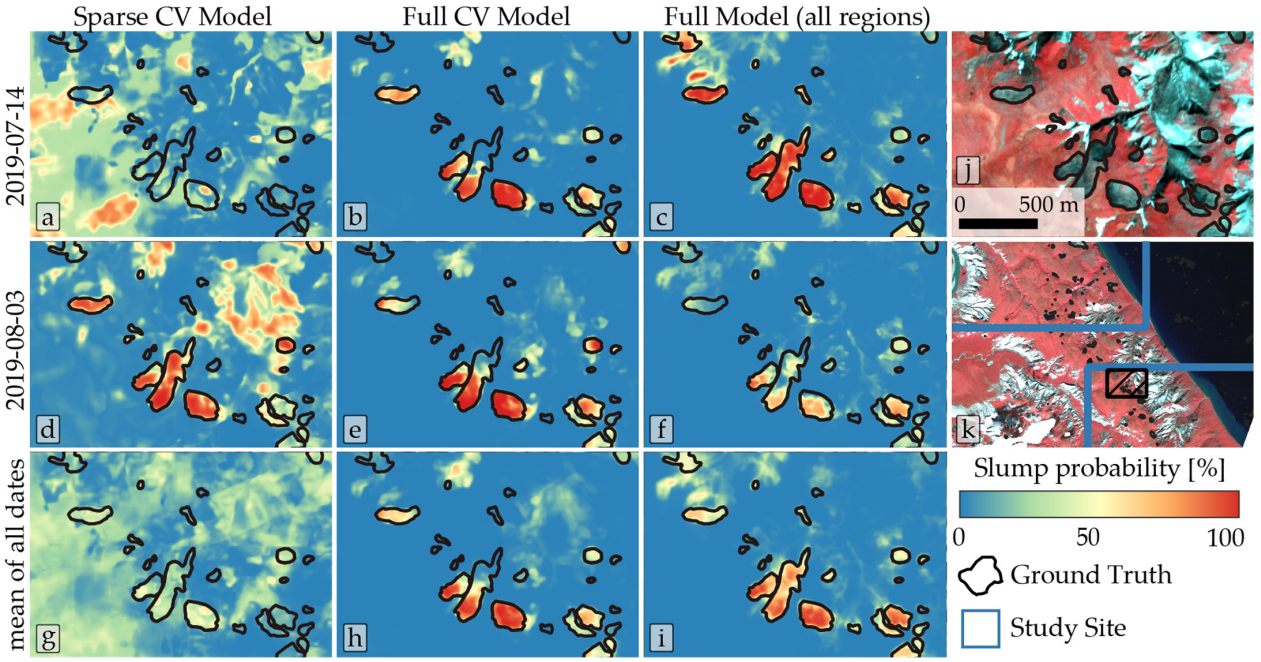


Figure 7. Detection results in the subset of the Horton Delta 02 study site with the modeled RTS probability on two different image dates ((**a**–**c**): 14 July 2019; (**d**–**f**): 03 August 2019) and the mean of all dates (**g**–**i**) as well as three different models ((**a**,**d**,**g**): sparse cross-validation model; (**b**,**e**,**h**): full cross-validation model; (**c**,**f**,**i**): full model trained on all regions). The subset of a multispectral false-color PlanetScope image (NIR-R-G) with ground truth is shown in panel (**j**). Panel (**k**) shows the approximate location of the subset within the study region.

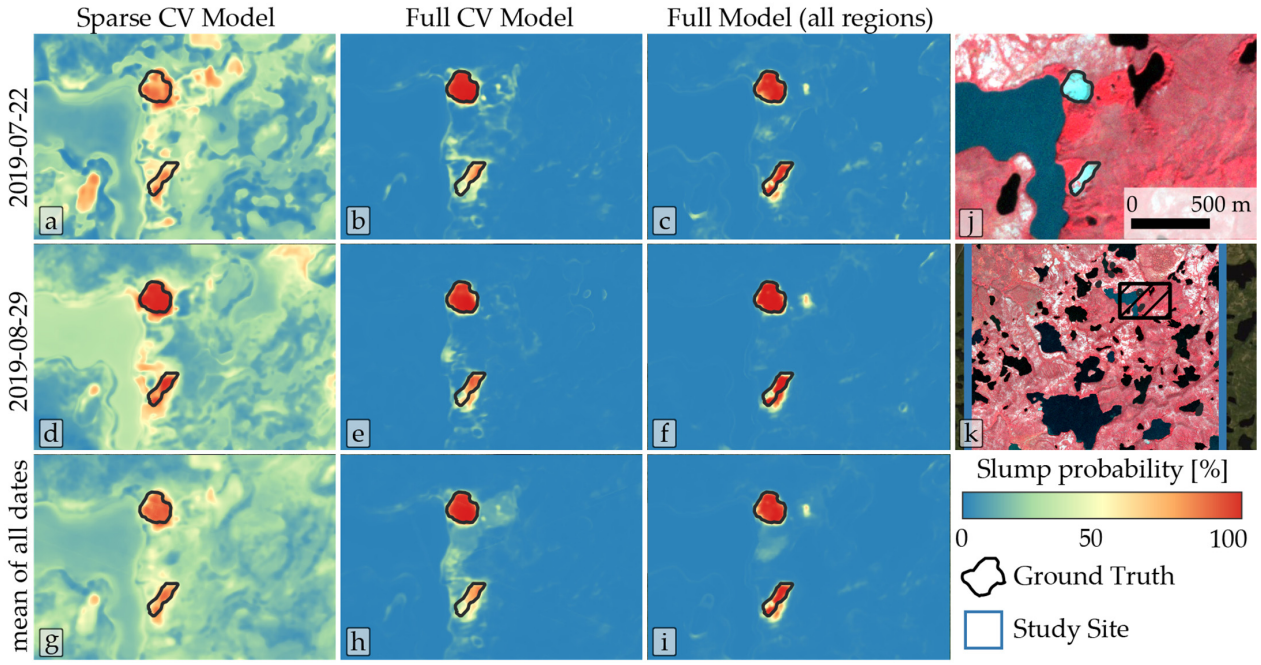


Figure 8. Detection results in the subset of the Lena River study site with the modeled RTS probability on two different image dates ((**a**–**c**): 22 July 2019; (**d**–**f**): 29 August 2019) and the mean of all dates (**g**–**i**) as well as three different models ((**a**,**d**,**g**): sparse cross-validation model; (**b**,**e**,**h**): full cross-validation model; (**c**,**f**,**i**): full model trained on all regions). The subset of a multispectral false-color PlanetScope image (NIR-R-G) with ground truth is shown in panel (**j**). Panel (**k**) shows the approximate location of the subset within the study region.

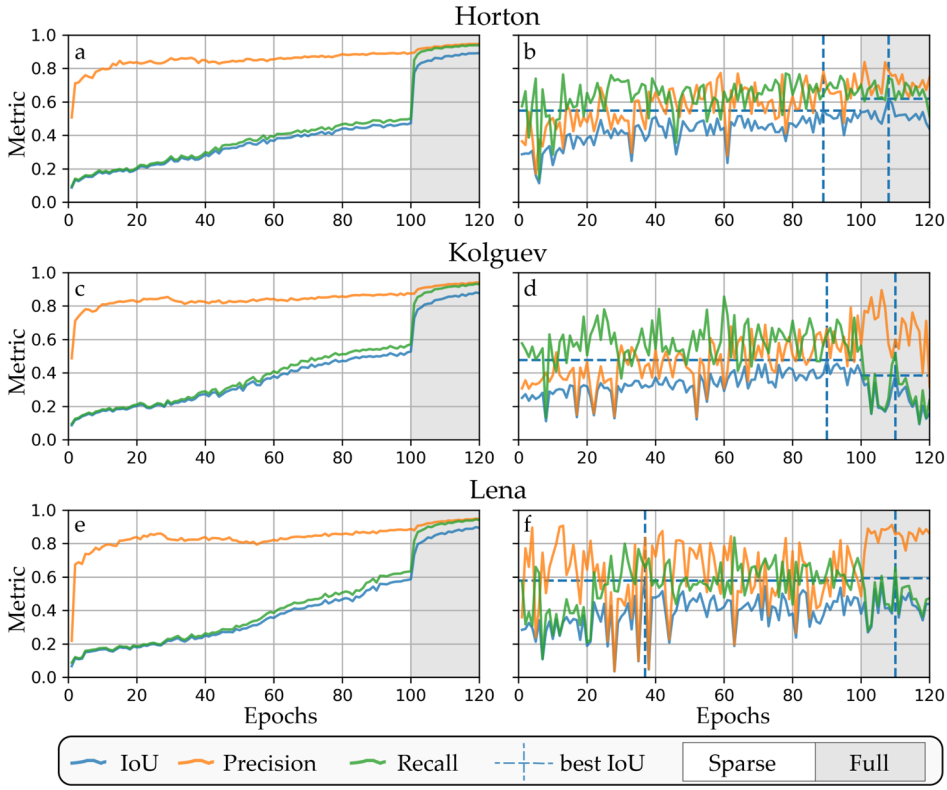


Figure 9. Training (**a**,**c**,**e**) and validation (**b**,**d**,**f**) metrics of best models for the three best regions with sparse and full training.

Non-slump/disturbed areas were closer to 0% probability, while thaw slumps typically showed *p*-values close to 100%. The stability of classifications was significantly improved after the full training, as seen in Figures 6–8, with comparable results between different dates (e.g., July and August).

However, misclassifications still occurred. False-positives occurred in rugged non-vegetated terrain (see Figures 6b,e,h and 7b,e,h) or silty water bodies (see Figure 8b,e,h). As these false-positives are inconsistent between different images dates, taking into account multiple images dates can help to detect and minimize false-positive objects (see Figure 8 bottom row).

False-negatives are prevalent in many classified datasets. In most cases, parts of thaw slumps were not detected. As seen in Figures 6–8, the slump area in proximity to the headwall was detected, whereas the distal parts remained undetected. This behavior suggests that the model is rather sensitive to the presence of headwall and thus steep slopes.

(2) The models trained on the full dataset, including the analyzed area, e.g., Horton (Figures 6c,f,i and 7c,f,i) or Lena (Figure 8c,f,i), performed well. When the model was trained on these datasets, the performance was high, as expected. The model also classified well when used for periods (2020, 2021) outside of the training data period (2018, 2019). Furthermore, RTS just outside the specific 10 × 10 km training sites, which were unknown to the model, were successfully identified.

The models also detected features that we did not define as RTS, but which have a similar appearance in remote sensing imagery. These are, e.g., borderline cases, where the distinction of slumps vs. non-slumps was difficult during the digitization processes, or other vegetation-less land surface types appeared. This further highlights the difficulties of manual thaw slump annotation/classification.

4. Discussion

The presented methodology provides a highly automated and reproducible proof of concept for the application of the novel deep learning-based segmentation of retrogressive thaw slumps across Arctic permafrost regions.

The results are promising, showing good agreement for some regions, with IoU scores of 0.55 and 0.58 for the best configurations. However, the performances for some of the regions, e.g., Tuktoyaktuk or Banks Island, were unsatisfactory and likely prone to overfitting. The comparison of model performance here to other studies and methods is hardly possible due to the different input data and regions and the lack of standardized training datasets. Still, many studies depend on manual or at least semi-automated methods [18,21] for detecting and segmenting RTS. Only Huang et al. [51] used a very similar deep learning methodology in the Beiluhe Region on the Tibetan Plateau. They achieved cross-validated F-scores of ~0.85, higher than our analysis with F-scores of 0.25 to 0.73. However, Huang et al. applied cross-validation within a single comparably homogeneous region, in contrast to the regional cross-validation approach across strongly varying landscapes in our study. High training accuracies and visual inference tests suggest a good model performance at least in proximity to the trained regions.

We tested different architectures, including UNet++, UNet, and DeepLabv3. The different model architectures performed similarly, but UNet++ produced on average the best results compared to UNet and DepLabV3, as shown in the original UNet++ paper [80].

The choice of encoders influenced the results only slightly, but on average, simpler encoders (Resnet34 > Resnet50 > Resnet101) achieved slightly better performances, although the original paper achieved higher accuracies with the more complex version [75]. We hypothesize that a simpler network might be slightly more resilient to overfitting. With a higher quantity and variability of training data across an even broader spatial extent, more complex and deeper architectures may become more favorable for segmenting RTS. As the technology is constantly evolving, with new DL architectures, packages and hardware, there is the potential for much further improvement in the near future.

The large range in the model performance between study sites compared to the performance between different model parameters suggests that regional landscape differences are by far the most influential factor in the successful delineation of RTS across permafrost landscapes. This magnifies the pressing need for representative and large training/ground truth datasets for specific geospatial targets, such as RTS in this case. Such a database does not exist yet for permafrost-specific features, in contrast to general remote sensing-based targets such as *PatternNet* [81] or *EuroSat* [82], or the standard photography databases, such as *ImageNet* [83]. Sufficiently large and spatially extensive ground truth data are particularly hard to find. The *ArcticNet* database [84] is the first remote sensing image database with a spatial focus on the Arctic, but this is limited to wetlands. For RTS, most openly accessible high-resolution polygon datasets are available for NW Canada [17,57], Alaska [25] and China [51,85]. For other studies, only RTS centroid coordinates are often made available in public archives [16], or detailed data are not accessible. Therefore, we want to propose the creation of an openly accessible pan-Arctic database for RTS and other important permafrost landscape features for the training of future DL-based applications aimed at detecting permafrost features and landscape change due to thaw and erosion.

However, such a database requires consistent data quality and standard procedures. During our manual ground truth creation, we encountered severe difficulty in delineating RTS. While the headwall was often clearly visible, the lower part of RTS was often highly ambiguous and hardly discernible. This difficulty makes the creation of consistent datasets, across different spatial regions and teams, even more challenging, thus requiring standardized protocols and a common effort among researchers.

The workflow is openly available (see code) and highly automated, and the data processing approach is transferable and reusable. However, access to VHR input data is required, which are largely only commercially available and/or accessible under very restricted licensing rules at this stage. This is a major limitation in transferability and scalability at the moment. Recently, Planet data are becoming more and more accessible to large groups of researchers free of charge through government-funded research programs, which allows their broader application in Big Data AI test cases such as our study.

The requirement for sufficiently powered hardware is very important. However, with the increasing level of GPU processing capacities, either in institutional systems or even freely accessible cloud services (e.g., Google colab), barriers against using AI-based systems will become increasingly lower for geoscientific object detection purposes.

The presented methodology has the potential to be used on a much larger spatial scale. However, such scaling to large regions requires more training data across different regions and better access to Planet data. Alternatively, free data sources, such as Sentinel-2, might be used as alternatives, but are limited by their lower spatial resolution used for small- to medium-sized landscape features.

5. Conclusions

With our study, we have laid the foundation for using deep learning-based methods to detect and segment RTS across the Arctic. Using a highly automated workflow in conjunction with state-of-the art DL model architectures, we were able to create sufficiently good and transferable models for several regions, as proven by regional cross-validation. Regional models worked sufficiently well, but spatial transferability is still an issue for some regions. Additionally, the creation of training datasets proved to be highly challenging due to the difficulties in delineating RTS. For scaling DL-based segmentation models to the entire pan-Arctic region, we propose a common effort to create large and high-quality training datasets to train and benchmark RTS detection models. With rapidly growing hardware capabilities and expanding data availability, the automated mapping and segmentation of RTS and other permafrost-related landscape features may be realized soon in order to better understand and predict the impact of climate change in the permafrost region.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/rs13214294/s1, Table S1: PlanetScope image_ids used for ground truth collection as well as model training and validation.

Author Contributions: I.N. designed the study, carried out most experiments and led the manuscript writing. K.H. led the software development and wrote the methodology section. S.B. created the majority of training datasets. G.G. co-acquired project funding and was responsible for supervision. All authors participated in writing, reviewing and editing the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by the HGF AI-CORE project. Additional funding was provided by the ESA CCI+ Permafrost and NSF Permafrost Discovery Gateway projects (NSF Grants #2052107 and #1927872). The MWFK Brandenburg provided funding for high-performance computing infrastructure within the Potsdam Arctic Innovation Lab at AWI.

Data Availability Statement: The main processing code is available at https://github.com/initze/thaw-slump-segmentation. Training Polygons and additional processing code are available at https://github.com/initze/DL_RTS_Paper. Unfortunately we cannot make the commercial PlanetScope input data available.

Acknowledgments: We thank the AWI IT department and the DLR EOC IT department for supporting the IT infrastructure for the HGF AI CORE project. We thank the reviewers for their valuable comments, which helped to improve the quality of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Appendix A

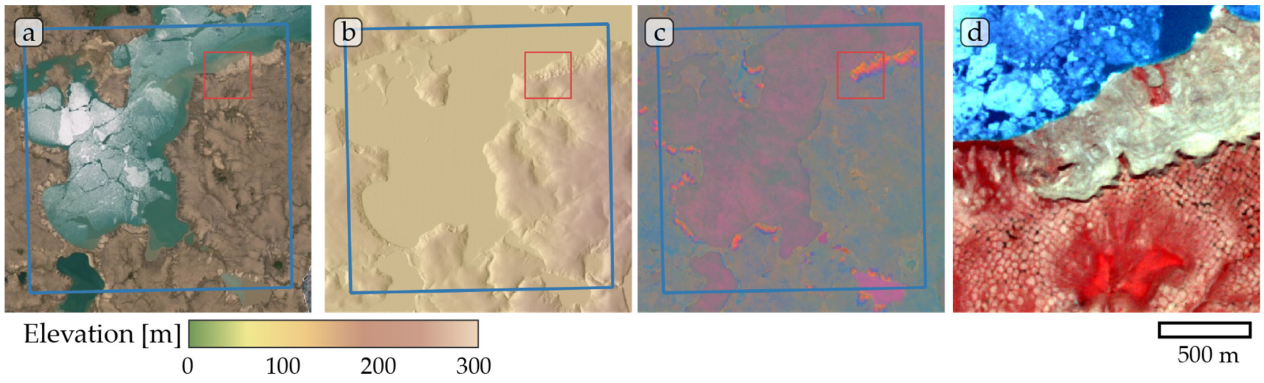


Figure A1. Study site Banks Island 01. (**a**) ESRI satellite layer, (**b**) ArcticDEM superimposed with hillshade, (**c**) Tasseled Cap trend visualization, (**d**) PlanetScope satellite image (NIR-R-G) acquired on 26 July 2019. Blue box, 10 × 10 km study site. Red box detailed view of (**d**).

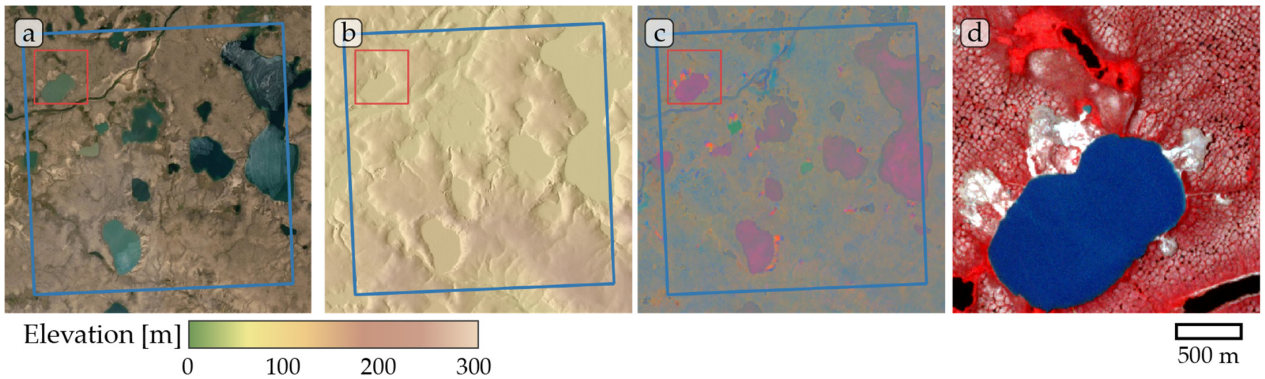


Figure A2. Study site Banks Island 02. (**a**) ESRI satellite layer, (**b**) ArcticDEM superimposed with hillshade, (**c**) Tasseled Cap trend visualization, (**d**) PlanetScope satellite image (NIR-R-G) acquired on 27 July 2019. Blue box, 10 × 10 km study site. Red box detailed view of (**d**).

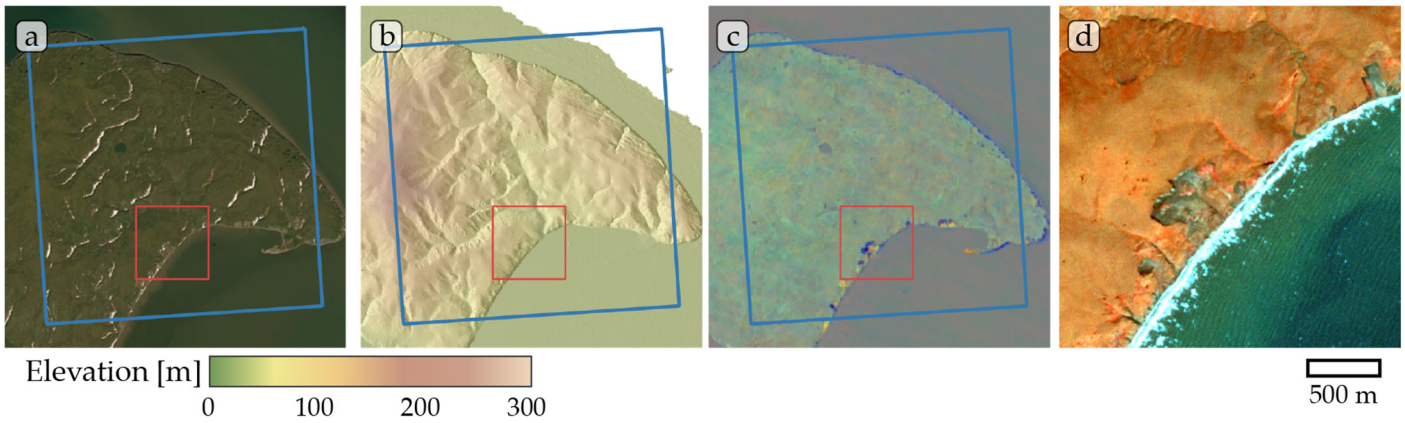


Figure A3. Study site Herschel Island 02. (**a**) ESRI satellite layer, (**b**) ArcticDEM superimposed with hillshade, (**c**) Tasseled Cap trend visualization, (**d**) PlanetScope satellite image (NIR-R-G) acquired on 02 September 2019. Blue box, 10 × 10 km study site. Red box detailed view of (**d**).

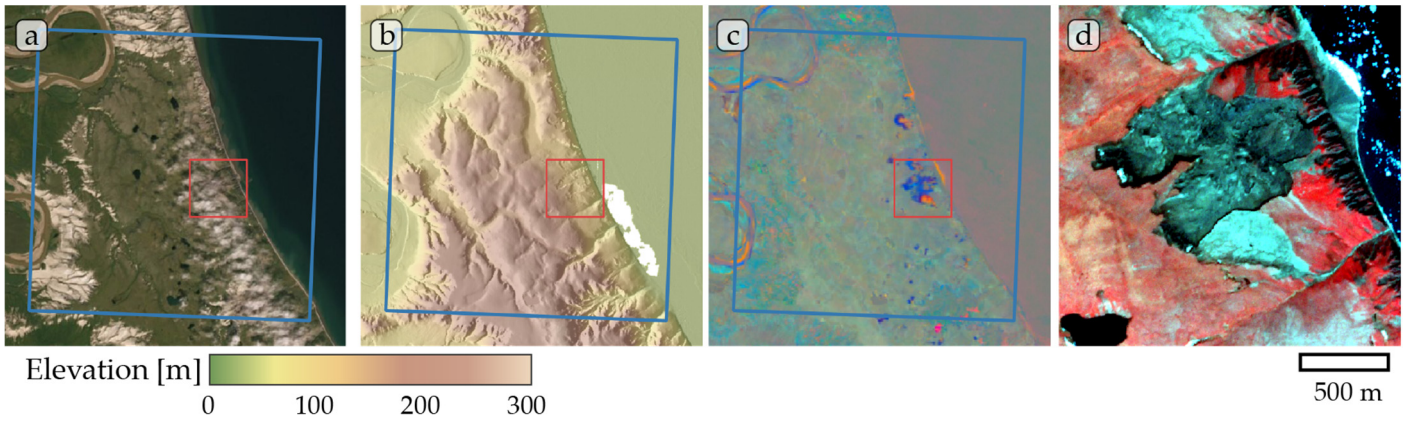


Figure A4. Study site Horton 01. (**a**) ESRI satellite layer, (**b**) ArcticDEM superimposed with hillshade, (**c**) Tasseled Cap trend visualization, (**d**) PlanetScope satellite image (NIR-R-G) acquired on 14 July 2019. Blue box, 10 × 10 km study site. Red box detailed view of (**d**).

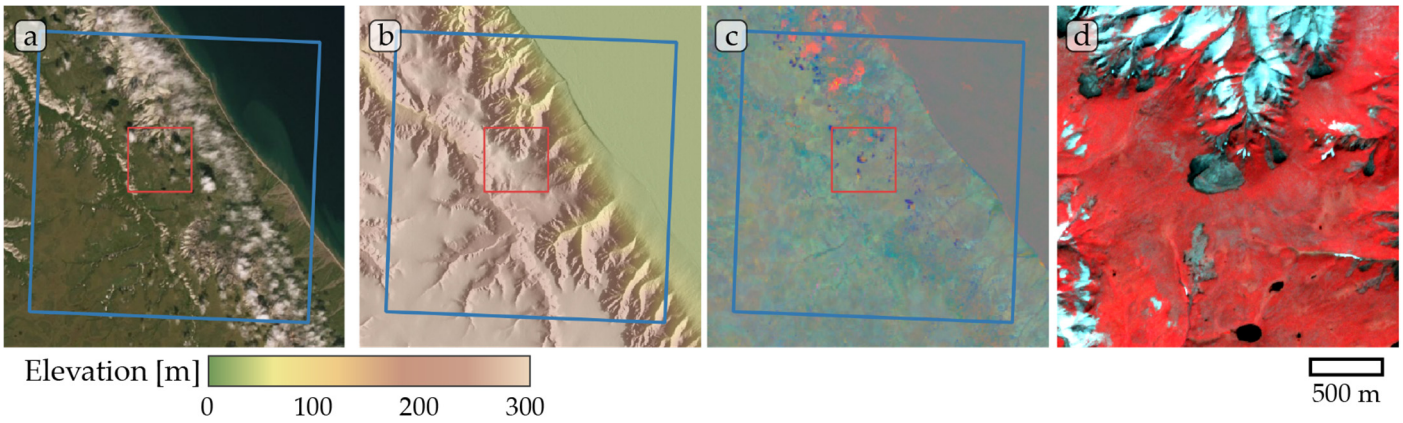


Figure A5. Study site Horton 02. (**a**) ESRI satellite layer, (**b**) ArcticDEM superimposed with hillshade, (**c**) Tasseled Cap trend visualization, (**d**) PlanetScope satellite image (NIR-R-G) acquired on 27 July 2019. Blue box, 10 × 10 km study site. Red box detailed view of (**d**).

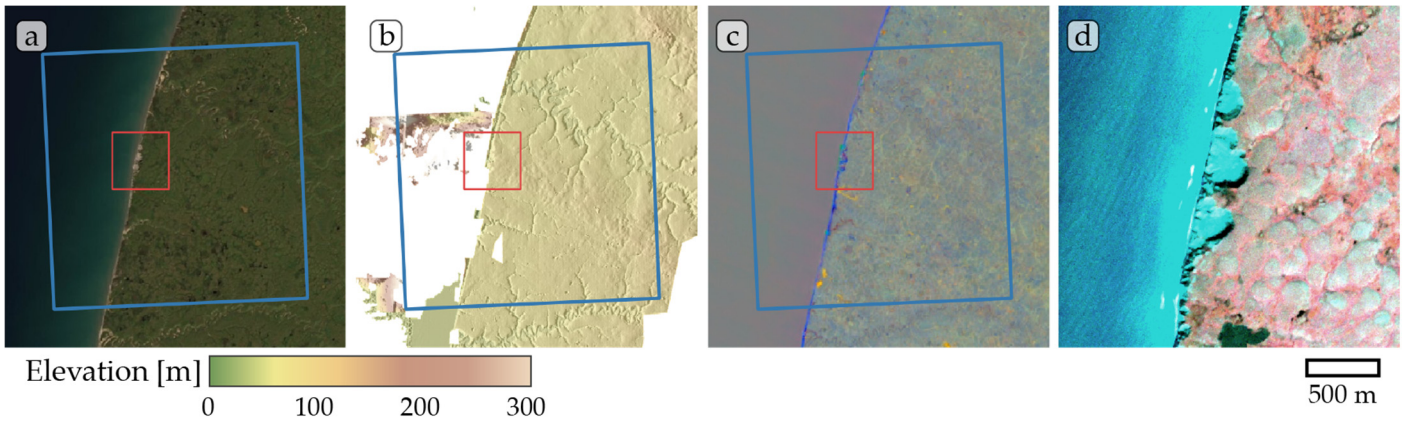


Figure A6. Study site Kolguev Island 01. (**a**) ESRI satellite layer, (**b**) ArcticDEM superimposed with hillshade, (**c**) Tasseled Cap trend visualization, (**d**) PlanetScope satellite image (NIR-R-G) acquired on 22 August 2019. Blue box, 10 × 10 km study site. Red box detailed view of (**d**). Note partially missing ArcticDEM data in (**b**).

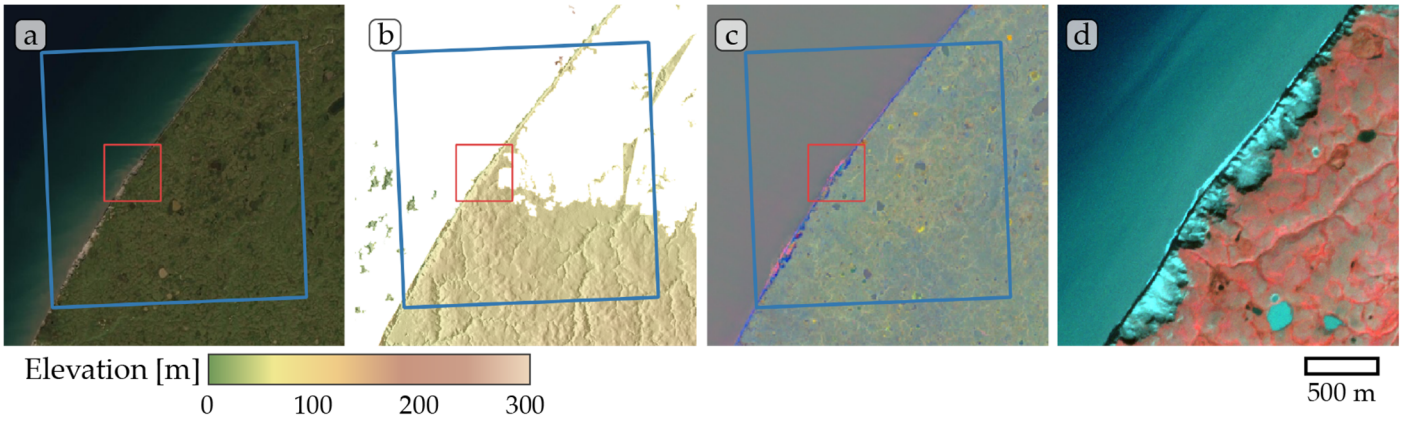


Figure A7. Study site Kolguev Island 02. (**a**) ESRI satellite layer, (**b**) ArcticDEM superimposed with hillshade, (**c**) Tasseled Cap trend visualization, (**d**) PlanetScope satellite image (NIR-R-G) acquired on 19 August 2019. Blue box, 10 × 10 km study site. Red box detailed view of (**d**). Note partially missing ArcticDEM data in (**b**).

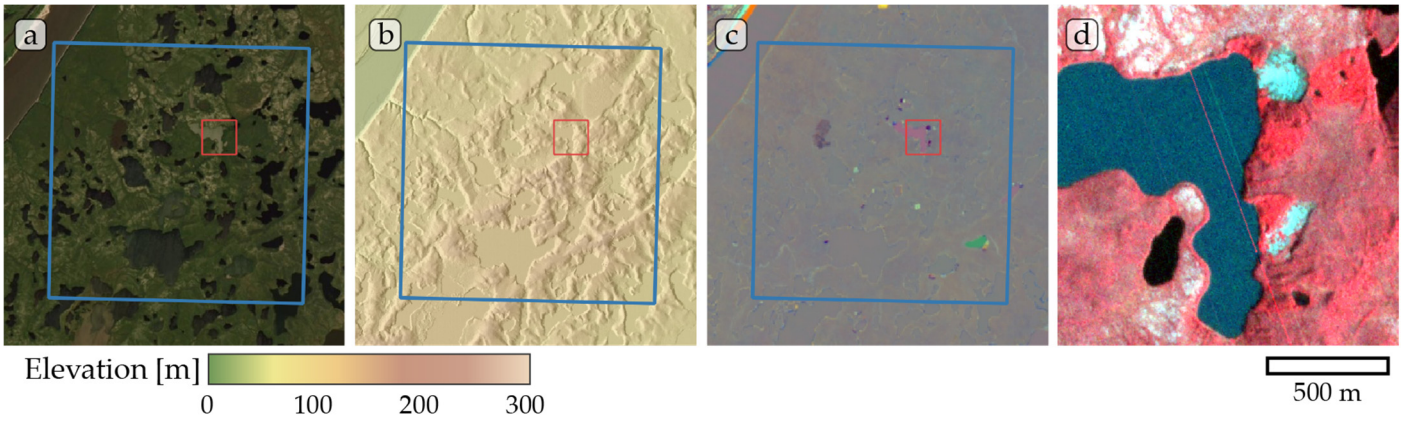


Figure A8. Study site Lena. (**a**) ESRI satellite layer, (**b**) ArcticDEM superimposed with hillshade, (**c**) Tasseled Cap trend visualization, (**d**) PlanetScope satellite image (NIR-R-G) acquired on 31 July 2019. Blue box, 10 × 10 km study site. Red box detailed view of (**d**).

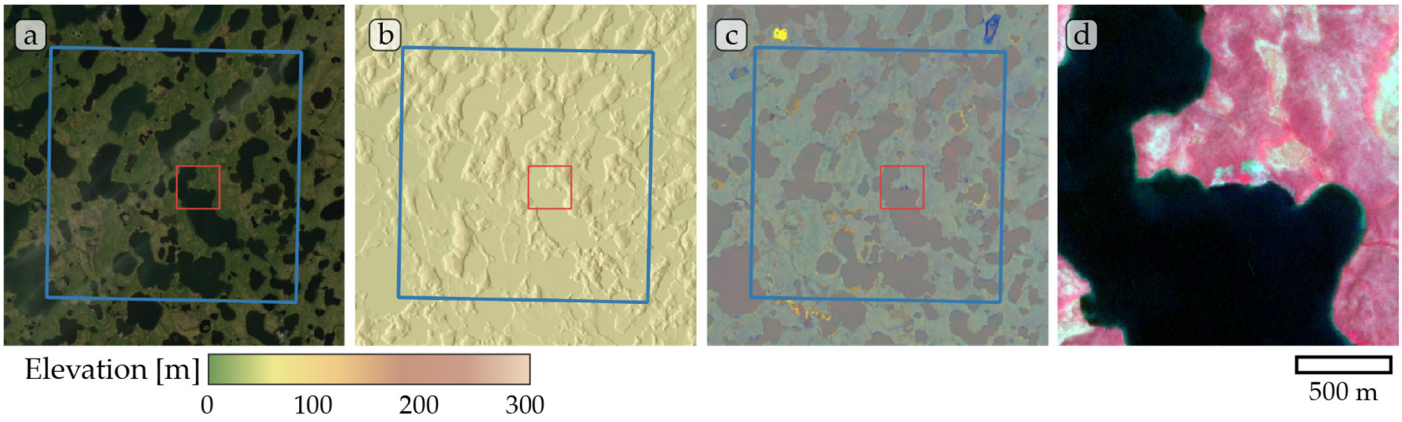


Figure A9. Study site Tuktoyaktuk. (**a**) ESRI satellite layer, (**b**) ArcticDEM superimposed with hillshade, (**c**) Tasseled Cap trend visualization, (**d**) PlanetScope satellite image (NIR-R-G) acquired on 13 July 2019. Blue box, 10 × 10 km study site. Red box detailed view of (**d**).

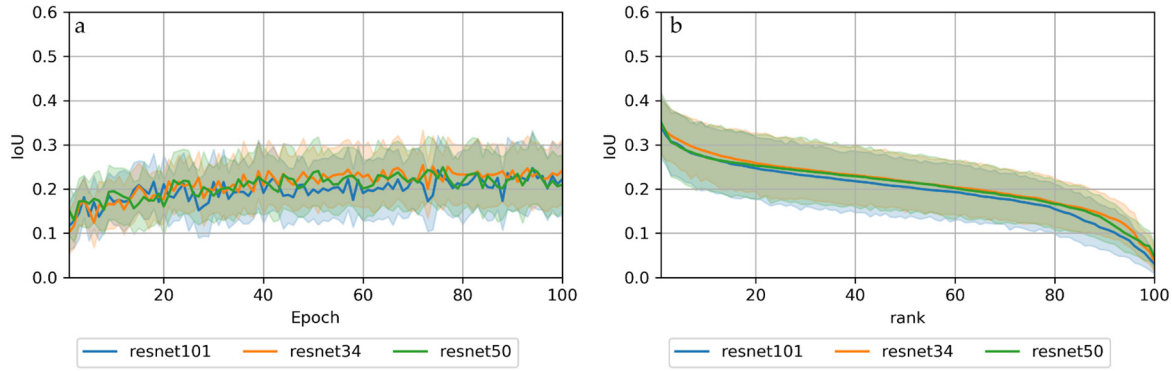


Figure A10. Mean and standard deviation of IoU scores per encoder sorted by (**a**) epoch and by (**b**) scores (best to worst). n = 18 (6 sites × 3 architectures) per encoder.

References

1. Box, J.E.; Colgan, W.T.; Christensen, T.R.; Schmidt, N.M.; Lund, M.; Parmentier, F.-J.W.; Brown, R.; Bhatt, U.S.; Euskirchen, E.S.; Romanovsky, V.E.; et al. Key Indicators of Arctic Climate Change: 1971–2017. *Environ. Res. Lett.* **2019**, *14*, 045010. [CrossRef]
2. Meredith, M.; Sommerkorn, M.; Cassotta, S.; Derksen, C.; Ekaykin, A.; Hollowed, A.; Kofinas, G.; Mackintosh, A.; Melbourne-Thomas, J.; Muelbert, M.M.C. *Polar Regions. Chapter 3, IPCC Special Report on the Ocean and Cryosphere in a Changing Climate*; WMO: Geneva, Switzerland, 2019.
3. Biskaborn, B.K.; Smith, S.L.; Noetzli, J.; Matthes, H.; Vieira, G.; Streletskiy, D.A.; Schoeneich, P.; Romanovsky, V.E.; Lewkowicz, A.G.; Abramov, A.; et al. Permafrost Is Warming at a Global Scale. *Nat. Commun.* **2019**, *10*, 264. [CrossRef]
4. Nitzbon, J.; Westermann, S.; Langer, M.; Martin, L.C.P.; Strauss, J.; Laboor, S.; Boike, J. Fast Response of Cold Ice-Rich Permafrost in Northeast Siberia to a Warming Climate. *Nat. Commun.* **2020**, *11*, 2201. [CrossRef] [PubMed]
5. Vasiliev, A.A.; Drozdov, D.S.; Gravis, A.G.; Malkova, G.V.; Nyland, K.E.; Streletskiy, D.A. Permafrost Degradation in the Western Russian Arctic. *Environ. Res. Lett.* **2020**, *15*, 045001. [CrossRef]
6. Hugelius, G.; Strauss, J.; Zubrzycki, S.; Harden, J.W.; Schuur, E.A.G.; Ping, C.-L.; Schirrmeister, L.; Grosse, G.; Michaelson, G.J.; Koven, C.D.; et al. Estimated Stocks of Circumpolar Permafrost Carbon with Quantified Uncertainty Ranges and Identified Data Gaps. *Biogeosciences* **2014**, *11*, 6573–6593. [CrossRef]
7. Strauss, J.; Schirrmeister, L.; Grosse, G.; Wetterich, S.; Ulrich, M.; Herzschuh, U.; Hubberten, H. The Deep Permafrost Carbon Pool of the Yedoma Region in Siberia and Alaska. *Geophys. Res. Lett.* **2013**, *40*, 6165–6170. [CrossRef] [PubMed]
8. Schuur, E.A.G.; McGuire, A.D.; Schädel, C.; Grosse, G.; Harden, J.W.; Hayes, D.J.; Hugelius, G.; Koven, C.D.; Kuhry, P.; Lawrence, D.M.; et al. Climate Change and the Permafrost Carbon Feedback. *Nature* **2015**, *520*, 171–179. [CrossRef]
9. Grosse, G.; Harden, J.; Turetsky, M.; McGuire, A.D.; Camill, P.; Tarnocai, C.; Frolking, S.; Schuur, E.A.G.; Jorgenson, T.; Marchenko, S.; et al. Vulnerability of High-Latitude Soil Organic Carbon in North America to Disturbance. *J. Geophys. Res. Biogeosci.* **2011**, *116*. [CrossRef]
10. Burn, C.R.; Lewkowicz, A.G. CANADIAN LANDFORM EXAMPLES - 17 RETROGRESSIVE THAW SLUMPS. *Can. Geogr. Géographe Can.* **1990**, *34*, 273–276. [CrossRef]

11. Lacelle, D.; Bjornson, J.; Lauriol, B. Climatic and Geomorphic Factors Affecting Contemporary (1950–2004) Activity of Retrogressive Thaw Slumps on the Aklavik Plateau, Richardson Mountains, NWT, Canada: Climatic and Geomorphic Factors Affecting Thaw Slump Activity. *Permafr. Periglac. Process.* **2010**, *21*, 1–15. [CrossRef]
12. Lantuit, H.; Pollard, W.H. Fifty Years of Coastal Erosion and Retrogressive Thaw Slump Activity on Herschel Island, Southern Beaufort Sea, Yukon Territory, Canada. *Geomorphology* **2008**, *95*, 84–102. [CrossRef]
13. Leibman, M.; Khomutov, A.; Kizyakov, A. Cryogenic Landslides in the West-Siberian Plain of Russia: Classification, Mechanisms, and Landforms. In *Landslides in Cold Regions in the Context of Climate Change*; Shan, W., Guo, Y., Wang, F., Marui, H., Strom, A., Eds.; Environmental Science and Engineering; Springer International Publishing: Cham, Germany, 2014; pp. 143–162. ISBN 978-3-319-00867-7.
14. Ashastina, K.; Schirrmeister, L.; Fuchs, M.; Kienast, F. Palaeoclimate Characteristics in Interior Siberia of MIS 6–2: First Insights from the Batagay Permafrost Mega-Thaw Slump in the Yana Highlands. *Clim. Past* **2017**, *13*, 795–818. [CrossRef]
15. Lantz, T.C.; Kokelj, S.V. Increasing Rates of Retrogressive Thaw Slump Activity in the Mackenzie Delta Region, N.W.T., Canada. *Geophys. Res. Lett.* **2008**, *35*, L06502. [CrossRef]
16. Lewkowicz, A.G.; Way, R.G. Extremes of Summer Climate Trigger Thousands of Thermokarst Landslides in a High Arctic Environment. *Nat. Commun.* **2019**, *10*, 1329. [CrossRef] [PubMed]
17. Segal, R.A.; Lantz, T.C.; Kokelj, S.V. Acceleration of Thaw Slump Activity in Glaciated Landscapes of the Western Canadian Arctic. *Environ. Res. Lett.* **2016**, *11*, 034025. [CrossRef]
18. Ward Jones, M.K.; Pollard, W.H.; Jones, B.M. Rapid Initialization of Retrogressive Thaw Slumps in the Canadian High Arctic and Their Response to Climate and Terrain Factors. *Environ. Res. Lett.* **2019**, *14*, 055006. [CrossRef]
19. Kokelj, S.V.; Lantz, T.C.; Kanigan, J.; Smith, S.L.; Coutts, R. Origin and Polycyclic Behaviour of Tundra Thaw Slumps, Mackenzie Delta Region, Northwest Territories, Canada. *Permafr. Periglac. Process.* **2009**, *20*, 173–184. [CrossRef]
20. Balser, A.W.; Jones, J.B.; Gens, R. Timing of Retrogressive Thaw Slump Initiation in the Noatak Basin, Northwest Alaska, USA. *J. Geophys. Res. Earth Surf.* **2014**, *119*, 1106–1120. [CrossRef]
21. Kokelj, S.V.; Lantz, T.C.; Tunnicliffe, J.; Segal, R.; Lacelle, D. Climate-Driven Thaw of Permafrost Preserved Glacial Landscapes, Northwestern Canada. *Geology* **2017**, *45*, 371–374. [CrossRef]
22. Andreev, A.A.; Grosse, G.; Schirrmeister, L.; Kuznetsova, T.V.; Kuzmina, S.A.; Bobrov, A.A.; Tarasov, P.E.; Novenko, E.Y.; Meyer, H.; Derevyagin, A.Y.; et al. Weichselian and Holocene Palaeoenvironmental History of the Bol'shoy Lyakhovsky Island, New Siberian Archipelago, Arctic Siberia. *Boreas* **2009**, *38*, 72–110. [CrossRef]
23. Lantuit, H.; Atkinson, D.; Paul Overduin, P.; Grigoriev, M.; Rachold, V.; Grosse, G.; Hubberten, H.-W. Coastal Erosion Dynamics on the Permafrost-Dominated Bykovsky Peninsula, North Siberia, 1951–2006. *Polar Res.* **2011**, *30*, 7341. [CrossRef]
24. Kokelj, S.V.; Tunnicliffe, J.; Lacelle, D.; Lantz, T.C.; Chin, K.S.; Fraser, R. Increased Precipitation Drives Mega Slump Development and Destabilization of Ice-Rich Permafrost Terrain, Northwestern Canada. *Glob. Planet. Chang.* **2015**, *129*, 56–68. [CrossRef]
25. Swanson, D.K. Permafrost Thaw-related Slope Failures in Alaska's Arctic National Parks. 1980–2019. *Permafr. Periglac. Process.* **2021**, *32*, 392–406. [CrossRef]
26. Babkina, E.A.; Leibman, M.O.; Dvornikov, Y.A.; Fakashchuk, N.Y.; Khairullin, R.R.; Khomutov, A.V. Activation of Cryogenic Processes in Central Yamal as a Result of Regional and Local Change in Climate and Thermal State of Permafrost. *Russ. Meteorol. Hydrol.* **2019**, *44*, 283–290. [CrossRef]
27. Kizyakov, A.; Zimin, M.; Leibman, M.; Pravikova, N. Monitoring of the Rate of Thermal Denudation and Thermal Abrasion on the Western Coast of Kolguev Island, Using High Resolution Satellite Images. *Kriosf. Zemli* **2013**, *17*, 36–47.
28. Vadakkedath, V.; Zawadzki, J.; Przeździecki, K. Multisensory Satellite Observations of the Expansion of the Batagaika Crater and Succession of Vegetation in Its Interior from 1991 to 2018. *Environ. Earth Sci.* **2020**, *79*, 150. [CrossRef]
29. Obu, J.; Lantuit, H.; Grosse, G.; Günther, F.; Sachs, T.; Helm, V.; Fritz, M. Coastal Erosion and Mass Wasting along the Canadian Beaufort Sea Based on Annual Airborne LiDAR Elevation Data. *Geomorphology* **2017**, *293*, 331–346. [CrossRef]
30. Swanson, D.; Nolan, M. Growth of Retrogressive Thaw Slumps in the Noatak Valley, Alaska, 2010–2016, Measured by Airborne Photogrammetry. *Remote Sens.* **2018**, *10*, 983. [CrossRef]
31. van der Sluijs, J.; Kokelj, S.; Fraser, R.; Tunnicliffe, J.; Lacelle, D. Permafrost Terrain Dynamics and Infrastructure Impacts Revealed by UAV Photogrammetry and Thermal Imaging. *Remote Sens.* **2018**, *10*, 1734. [CrossRef]
32. Nitze, I.; Grosse, G.; Jones, B.M.; Romanovsky, V.E.; Boike, J. Remote Sensing Quantifies Widespread Abundance of Permafrost Region Disturbances across the Arctic and Subarctic. *Nat. Commun.* **2018**, *9*, 5423. [CrossRef] [PubMed]
33. Lara, M.J.; Chipman, M.L.; Hu, F.S. Automated Detection of Thermoerosion in Permafrost Ecosystems Using Temporally Dense Landsat Image Stacks. *Remote Sens. Environ.* **2019**, *221*, 462–473. [CrossRef]
34. Runge, A.; Nitze, I.; Grosse, G. Remote Sensing Annual Dynamics of Rapid Permafrost Thaw Disturbances with LandTrendr. *Remote Sens. Environ.*. accepted.
35. Blaschke, T.; Lang, S.; Hay, G. *Object-Based Image Analysis: Spatial Concepts for Knowledge-Driven Remote Sensing Applications*; Springer Science & Business Media: Berlin/Heidelberg Germany, 2008.
36. Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.-S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–36. [CrossRef]
37. Ghorbanzadeh, O.; Blaschke, T.; Gholamnia, K.; Meena, S.; Tiede, D.; Aryal, J. Evaluation of Different Machine Learning Methods and Deep-Learning Convolutional Neural Networks for Landslide Detection. *Remote Sens.* **2019**, *11*, 196. [CrossRef]

38. Prakash, N.; Manconi, A.; Loew, S. A New Strategy to Map Landslides with a Generalized Convolutional Neural Network. *Sci. Rep.* **2021**, *11*, 9722. [CrossRef] [PubMed]
39. Wang, H.; Zhang, L.; Yin, K.; Luo, H.; Li, J. Landslide Identification Using Machine Learning. *Geosci. Front.* **2021**, *12*, 351–364. [CrossRef]
40. Wagner, F.; Dalagnol, R.; Tarabalka, Y.; Segantine, T.; Thomé, R.; Hirye, M. U-Net-Id, an Instance Segmentation Model for Building Extraction from Satellite Images—Case Study in the Joanópolis City, Brazil. *Remote Sens.* **2020**, *12*, 1544. [CrossRef]
41. Yang, N.; Tang, H. GeoBoost: An Incremental Deep Learning Approach toward Global Mapping of Buildings from VHR Remote Sensing Images. *Remote Sens.* **2020**, *12*, 1794. [CrossRef]
42. Zhao, W.; Du, S.; Emery, W.J. Object-Based Convolutional Neural Network for High-Resolution Imagery Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *10*, 3386–3396. [CrossRef]
43. Wagner, F.H.; Dalagnol, R.; Tagle Casapia, X.; Streher, A.S.; Phillips, O.L.; Gloor, E.; Aragão, L.E.O.C. Regional Mapping and Spatial Distribution Analysis of Canopy Palms in an Amazon Forest Using Deep Learning and VHR Images. *Remote Sens.* **2020**, *12*, 2225. [CrossRef]
44. Abdalla, A.; Cen, H.; Abdel-Rahman, E.; Wan, L.; He, Y. Color Calibration of Proximal Sensing RGB Images of Oilseed Rape Canopy via Deep Learning Combined with K-Means Algorithm. *Remote Sens.* **2019**, *11*, 3001. [CrossRef]
45. Bhuiyan, M.A.E.; Witharana, C.; Liljedahl, A.K.; Jones, B.M.; Daanen, R.; Epstein, H.E.; Kent, K.; Griffin, C.G.; Agnew, A. Understanding the Effects of Optimal Combination of Spectral Bands on Deep Learning Model Predictions: A Case Study Based on Permafrost Tundra Landform Mapping Using High Resolution Multispectral Satellite Imagery. *J. Imaging* **2020**, *6*, 97. [CrossRef] [PubMed]
46. Park, J.H.; Inamori, T.; Hamaguchi, R.; Otsuki, K.; Kim, J.E.; Yamaoka, K. RGB Image Prioritization Using Convolutional Neural Network on a Microprocessor for Nanosatellites. *Remote Sens.* **2020**, *12*, 3941. [CrossRef]
47. Abolt, C.J.; Young, M.H.; Atchley, A.L.; Wilson, C.J. Brief Communication: Rapid Machine-Learning-Based Extraction and Measurement of Ice Wedge Polygons in High-Resolution Digital Elevation Models. *Cryosphere* **2019**, *13*, 237–245. [CrossRef]
48. Bhuiyan, M.A.E.; Witharana, C.; Liljedahl, A.K. Use of Very High Spatial Resolution Commercial Satellite Imagery and Deep Learning to Automatically Map Ice-Wedge Polygons across Tundra Vegetation Types. *J. Imaging* **2020**, *6*, 137. [CrossRef]
49. Zhang, W.; Liljedahl, A.K.; Kanevskiy, M.; Epstein, H.E.; Jones, B.M.; Jorgenson, M.T.; Kent, K. Transferability of the Deep Learning Mask R-CNN Model for Automated Mapping of Ice-Wedge Polygons in High-Resolution Satellite and UAV Images. *Remote Sens.* **2020**, *12*, 1085. [CrossRef]
50. Bartsch, A.; Pointner, G.; Ingeman-Nielsen, T.; Lu, W. Towards Circumpolar Mapping of Arctic Settlements and Infrastructure Based on Sentinel-1 and Sentinel-2. *Remote Sens.* **2020**, *12*, 2368. [CrossRef]
51. Huang, L.; Luo, J.; Lin, Z.; Niu, F.; Liu, L. Using Deep Learning to Map Retrogressive Thaw Slumps in the Beiluhe Region (Tibetan Plateau) from CubeSat Images. *Remote Sens. Environ.* **2020**, *237*, 111534. [CrossRef]
52. Huang, L.; Liu, L.; Luo, J.; Lin, Z.; Niu, F. Automatically Quantifying Evolution of Retrogressive Thaw Slumps in Beiluhe (Tibetan Plateau) from Multi-Temporal CubeSat Images. *Int. J. Appl. Earth Obs. Geoinform.* **2021**, *102*, 102399. [CrossRef]
53. Lakeman, T.R.; England, J.H. Paleoglaciological Insights from the Age and Morphology of the Jesse Moraine Belt, Western Canadian Arctic. *Quat. Sci. Rev.* **2012**, *47*, 82–100. [CrossRef]
54. Fraser, R.; Kokelj, S.; Lantz, T.; McFarlane-Winchester, M.; Olthof, I.; Lacelle, D. Climate Sensitivity of High Arctic Permafrost Terrain Demonstrated by Widespread Ice-Wedge Thermokarst on Banks Island. *Remote Sens.* **2018**, *10*, 954. [CrossRef]
55. Walker, D.A.; Raynolds, M.K.; Daniëls, F.J.A.; Einarsson, E.; Elvebakk, A.; Gould, W.A.; Katenin, A.E.; Kholod, S.S.; Markon, C.J.; Melnikov, E.S.; et al. The Circumpolar Arctic Vegetation Map. *J. Veg. Sci.* **2005**, *16*, 267–282. [CrossRef]
56. Obu, J.; Westermann, S.; Kääb, A.; Bartsch, A. *Ground Temperature Map, 2000-2016, Northern Hemisphere Permafrost in Earth & Environmental Science*; PANGAEA, 2018. [CrossRef]
57. Ramage, J.L.; Irrgang, A.M.; Herzschuh, U.; Morgenstern, A.; Couture, N.; Lantuit, H. Terrain Controls on the Occurrence of Coastal Retrogressive Thaw Slumps along the Yukon Coast, Canada: Coastal RTSs Along the Yukon Coast. *J. Geophys. Res. Earth Surf.* **2017**, *122*, 1619–1634. [CrossRef]
58. Fritz, M.; Wetterich, S.; Meyer, H.; Schirrmeister, L.; Lantuit, H.; Pollard, W.H. Origin and Characteristics of Massive Ground Ice on Herschel Island (Western Canadian Arctic) as Revealed by Stable Water Isotope and Hydrochemical Signatures: Origin and Characteristics of Massive Ground Ice on Herschel Island. *Permafr. Periglac. Process.* **2011**, *22*, 26–38. [CrossRef]
59. Stauch, G.; Lehmkuhl, F. Quaternary Glaciations in the Verkhoyansk Mountains, Northeast Siberia. *Quat. Res.* **2010**, *74*, 145–155. [CrossRef]
60. Burn, C.R.; Kokelj, S.V. The Environment and Permafrost of the Mackenzie Delta Area. *Permafr. Periglac. Process.* **2009**, *20*, 83–105. [CrossRef]
61. Olthof, I.; Fraser, R.H.; Schmitt, C. Landsat-Based Mapping of Thermokarst Lake Dynamics on the Tuktoyaktuk Coastal Plain, Northwest Territories, Canada since 1985. *Remote Sens. Environ.* **2015**, *168*, 194–204. [CrossRef]
62. Plug, L.J.; Walls, C.; Scott, B.M. Tundra Lake Changes from 1978 to 2001 on the Tuktoyaktuk Peninsula, Western Canadian Arctic. *Geophys. Res. Lett.* **2008**, *35*, L03502. [CrossRef]
63. Planet Team. Planet Application Program Interface: In Space for Life on Earth. Planet. 2017. Available online: https://api.planet.com (accessed on 16 July 2021).

64. Porter, C.; Morin, P.; Howat, I.; Noh, M.-J.; Bates, B.; Peterman, K.; Keesey, S.; Schlenk, M.; Gardiner, J.; Tomko, K.; et al. *ArcticDEM [Data set]*; Harvard Dataverse, 2018. [CrossRef]
65. Li, J.; Roy, D. A Global Analysis of Sentinel-2A, Sentinel-2B and Landsat-8 Data Revisit Intervals and Implications for Terrestrial Monitoring. *Remote Sens.* **2017**, *9*, 902. [CrossRef]
66. Roy, S.; Swetnam, T.L. *Tyson-Swetnam/Porder: Porder: Simple CLI for Planet OrdersV2 API*; Zenodo, 2020.
67. Nitze, I.; Grosse, G.; Jones, B.; Arp, C.; Ulrich, M.; Fedorov, A.; Veremeeva, A. Landsat-Based Trend Analysis of Lake Dynamics across Northern Permafrost Regions. *Remote Sens.* **2017**, *9*, 640. [CrossRef]
68. Nitze, I.; Grosse, G. Detection of Landscape Dynamics in the Arctic Lena Delta with Temporally Dense Landsat Time-Series Stacks. *Remote Sens. Environ.* **2016**, *181*, 27–41. [CrossRef]
69. Huang, C.; Wylie, B.; Yang, L.; Homer, C.; Zylstra, G. Derivation of a Tasselled Cap Transformation Based on Landsat 7 At-Satellite Reflectance. *Int. J. Remote Sens.* **2002**, *23*, 1741–1748. [CrossRef]
70. QGIS Development Team. *QGIS Geographic Information System*; QGIS Association, 2021.
71. Nextgis/Quickmapservices. 2021. [Python]. NextGIS. (Original Work Published 2014). Available online: https://github.com/nextgis/quickmapservices (accessed on 7 July 2021).
72. Segal, R.; Kokelj, S.; Lantz, T.; Durkee, K.; Gervais, S.; Mahon, E.; Snijders, M.; Buysse, J.; Schwarz, S. Broad-Scale Mapping of Terrain Impacted by Retrogressive Thaw Slumping in Northwestern Canada. *NWT Open Rep.* **2016**, *8*, 1–17.
73. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *ArXiv191201703 Cs Stat* **2019**.
74. Buslaev, A.; Iglovikov, V.I.; Khvedchenya, E.; Parinov, A.; Druzhinin, M.; Kalinin, A.A. Albumentations: Fast and Flexible Image Augmentations. *Information* **2020**, *11*, 125. [CrossRef]
75. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
76. Ronneberger, O. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention; Springer: Cham, Germany, 2015; pp. 234–241.
77. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Stoyanov, D., Taylor, Z., Carneiro, G., Syeda-Mahmood, T., Martel, A., Maier-Hein, L., Tavares, J.M.R.S., Bradley, A., Papa, J.P., Belagiannis, V., et al., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Germany, 2018; Volume 11045, pp. 3–11. ISBN 978-3-030-00888-8.
78. Chen, L.-C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *ArXiv170605587 Cs* **2017**.
79. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *ArXiv14126980 Cs* **2017**.
80. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. *IEEE Trans. Med. Imaging* **2020**, *39*, 1856–1867. [CrossRef]
81. Zhou, W.; Newsam, S.; Li, C.; Shao, Z. PatternNet: A Benchmark Dataset for Performance Evaluation of Remote Sensing Image Retrieval. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 197–209. [CrossRef]
82. Helber, P.; Bischke, B.; Dengel, A.; Borth, D. EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 2217–2226. [CrossRef]
83. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
84. Jiang, Z.; Von Ness, K.; Loisel, J.; Wang, Z. ArcticNet: A Deep Learning Solution to Classify the Arctic Area. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019; pp. 38–47.
85. Xia, Z.; Huang, L.; Liu, L. An Inventory of Retrogressive Thaw Slumps Along the Vulnerable Qinghai-Tibet Engineering Corridor. 2021. Available online: https://doi.pangaea.de/10.1594/PANGAEA.933957 (accessed on 19 October 2021).

A.5 PixelDINO: Semi-Supervised Semantic Segmentation for Detecting Permafrost Disturbances

Reference

K. Heidler, I. Nitze, G. Grosse, and X. X. Zhu, "PixelDINO: Semi-supervised semantic segmentation for detecting permafrost disturbances," *IEEE Transactions on Geoscience and Remote Sensing (in review)*, 2024. DOI: 10.48550/arXiv.2401.09271

Copyright

PixelDINO: Semi-Supervised Semantic Segmentation for Detecting Permafrost Disturbances in the Arctic

Konrad Heidler, *Student Member, IEEE*, Ingmar Nitze, Guido Grosse, and Xiao Xiang Zhu, *Fellow, IEEE*

***Abstract*—Arctic Permafrost is facing significant changes due to global climate change. As these regions are largely inaccessible, remote sensing plays a crucial rule in better understanding the underlying processes across the Arctic. In this study, we focus on the remote detection of Retrogressive Thaw Slumps (RTSs), a permafrost disturbance comparable to slow landslides. For such remote sensing tasks, deep learning has become an indispensable tool, but limited labeled training data remains a challenge for training accurate models. We present PixelDINO, a semi-supervised learning approach, to improve model generalization across the Arctic with a limited number of labels. PixelDINO leverages unlabeled data by training the model to define its own segmentation categories (pseudo-classes), promoting consistent structural learning across strong data augmentations. This allows the model to extract structural information from unlabeled data, supplementing the learning from labeled data. PixelDINO surpasses both supervised baselines and existing semi-supervised methods, achieving average Intersection-over-Union (IoU) of 30.2 and 39.5 on the two evaluation sets, representing significant improvements of 13% and 21%, respectively over the strongest existing models. This highlights the potential for training robust models that generalize well to regions that were not included in the training data.**

***Index Terms*—Semi-Supervised Learning, Semantic Segmentation, Permafrost, Retrogressive Thaw Slumps, Self-Distillation without Labels**

I. Introduction

In step with global climate change, permafrost is changing rapidly. Rising temperatures in the Arctic have large implications for perennially frozen soil which can destabilize upon the thawing of ice-rich ground. Owing to their remoteness and sparse population, permafrost areas are often difficult to access physically. Therefore, in-situ measurements are only available for specific study sites at specific dates when expeditions visited that site or when data is collected through local sensors [1]. Therefore, Remote sensing techniques are a valuable method that can monitor permafrost on a pan-Arctic scale, and a useful approach for upscaling and understanding of broad spatio-temporal dynamics of permafrost thaw processes [2], [3]. To further improve the efficiency of remote sensing monitoring for these applications, machine learning techniques offer great potential in automating laborious annotation tasks.

Permafrost is generally a subsurface phenomenon, making it difficult to observe from satellite observations. Other than permafrost itself, permafrost degradation landforms like retrogressive thaw slumps (RTSs) are visible in optical satellite imagery due to their distinct shape and spectral signature compared to the surrounding regions. This makes them a viable target of study via remote sensing methods. RTSs are mass movements akin to slow-flowing landslides caused by melting of massive ground-ice in permafrost regions. [4]. RTSs are rather small features generally measuring less than $10\,\mathrm{ha}$ in area [5], [6], with some notable exceptions, so-called megaslumps, exceeding $40\,\mathrm{ha}$ [7]. RTSs form due to specific local environmental conditions like slope, landscape history, ground temperature, and disturbances [4]. They typically occur in glacial moraines with preserved remnant glacial ice, syngenetic ice-rich yedoma permafrost, or marine deposits, which were raised due to isostatic uplift [8]. Understanding and quantifying RTS dynamics is important as they pose potential hazards to infrastructure [9], directly affect water quality in downstream aquatic environments [10], and locally mobilize large amounts of formerly frozen sediment and organic matter [8].

Machine learning, specifically deep learning, can automate the identification of RTSs from satellite imagery. Existing studies often achieve mixed results, which in many cases can be attributed to the algorithms' requirements for an extensive collection of labeled training data that is hard to acquire in large volumes [11]–[15]. While decent prediction results are obtained for selected study sites, accurate pan-Arctic generalization remains an elusive goal [12], [15].

This study explores how to make models better generalize to previously unseen regions. While increasing the available training data through additional labelling efforts is one option, it comes at a large labor cost for the involved domain experts. In an attempt to tackle this issue from a methodological angle instead, we explore semi-supervised learning for improving model performance without the need for additional annotated training data. In classical *supervised learning*, a model is trained on labeled data only. In contrast to this, *self-supervised learning* aims to train models without any labels. Combining

K. Heidler and X. Zhu are with the Chair of Data Science in Earth Observation (SiPEO), Department of Aerospace and Geodesy, School of Engineering and Design, Technical University of Munich (TUM), Munich, Germany. E-mails: k.heidler@tum.de; xiaoxiang.zhu@tum.de

X. Zhu is also with the Munich Center for Machine Learning, 80333, Munich, Germany.

I. Nitze and G. Grosse are with the Permafrost Research Section, Alfred Wegener Institute Helholtz Centre for Polar and Marine Research, Potsdam, Germany. E-mails: ingmar.nitze@awi.de, guido.grosse@awi.de

G. Grosse is also with the Institute of Geosciences, University of Potsdam, Potsdam, Germany.

This work was supported by the BMWK project ML4Earth. XZ was further supported by the BMBF future lab AI4EO and the Munich Center for Machine Learning (MCML). GG and IN were further supported by the projects HGF AI-CORE, and NSF Permafrost Discovery Gateway.

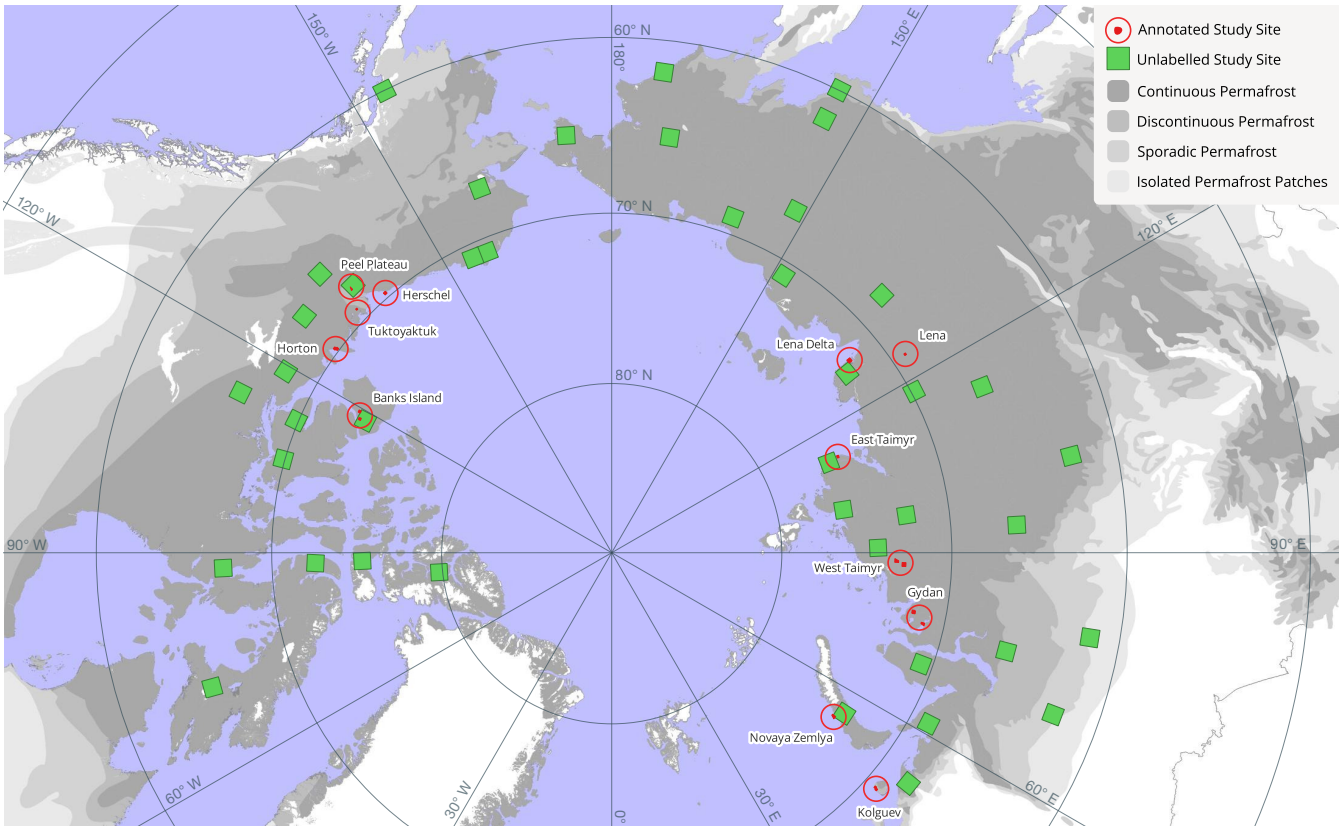


Fig. 1. Spatial distribution of the annotated training sites (red). It can be seen that the labeled data has quite limited spatial coverage. By using semi-supervised learning, it is possible to include large areas of unlabeled Sentinel-2 imagery (green) into the training process. Basemap source: [16]

TABLE I
STATISTICS FOR THE STUDY REGIONS (ORDERED BY LONGITUDE)

	RTS		Satellite Images	
Region	Count	Area [km^2]	Count	Area [km^2]
Herschel	148	1.6	10	442.9
Peel Plateau	37	0.68	1	87.9
Tuktoyaktuk	391	1.3	19	899.4
Horton	534	13.2	18	866.0
Banks Island	552	28.2	20	814.6
Kolguev	319	12.6	34	1814.1
Novaya Zemlya	982	12.3	3	454.0
Gydan	50	0.2	2	966.9
West Taimyr	110	0.5	2	1057.1
East Taimyr	839	9.2	3	148.9
Lena	238	4.2	41	2020.6
Lena Delta	136	0.8	1	625.5

these two paradigms, *semi-supervised learning* trains models on both labeled and unlabeled data at the same time [17], [18]. This strategy allows for the inclusion of unlabeled satellite imagery into the training process. While labelling is a laborious task, the underlying satellite imagery is openly available. Therefore, semi-supervised learning methods are exceptionally well-suited for remote sensing tasks.

In this study, we propose a new framework for semi-supervised semantic segmentation called *PixelDINO*. Our framework builds on the successful self-supervised learning framework DINO [19], which was originally developed to learn features for image classification. The main idea behind DINO is *self-distillation with no labels*, which is a special case of knowledge distillation. In *knowledge distillation*, a model is trained to closely match another model's outputs in order to transfer learned knowledge from one model to another. Self-distillation with no labels describes distillating a model's knowledge into itself while applying certain transformations to the data [19]. We adopt this idea to pixel-wise prediction tasks like semantic segmentation and then combine it with a regular supervised learning procedure into a semi-supervised learning framework.

As shown in Fig. 1, spatial coverage of the Arctic can be greatly improved for RTS detection by including unlabeled data in a semi-supervised fashion. Using this dataset, we present experimental results for the task of RTS detection, where we demonstrate that PixelDINO outperforms both supervised baseline methods and other semi-supervised semantic segmentation approaches.

II. Related Work

In order to place our contributions into a larger scientific context, this section summarizes existing research on monitoring RTSs with remote sensing, and gives an overview of representation learning and semi-supervised segmentation methods in remote sensing.

A. Monitoring Retrogressive Thaw Slumps

As permafrost cannot be directly seen from space, many permafrost remote sensing studies focus instead on monitoring specific targets that are known or assumed to be correlated with the state of permafrost or its vulnerability [3]. Spatially consistent monitoring of specific permafrost degradation landforms with high temporal resolution is a desirable goal, since it would allow assessments regarding vulnerability of local infrastructure and the biogeochemical implications of rapid permafrost thaw for both the local environment and the global climate system [12].

The detection of such features in satellite imagery is not without challenges. Retrogressive Thaw Slumps in permafrost regions are often hard to detect due to their widespread distribution, small size, and their varying stages of activity [12], [15]. Further, optical remote sensing is inhibited by snow cover, cloud cover, and polar night for large parts of the year, so that features can only be reliably detected during the summer months [3].

Regarding data sources, permafrost disturbances can be mapped using different remote sensing approaches, such as optical image analysis [12], optical time series analysis [20], surface elevation data [21] or interferometric synthetic aperture radar (InSAR) measurements [22].

Many studies rely on manual digitization of permafrost disturbance landforms in satellite imagery [23], [24]. While this approach ensures good accuracy, it quickly becomes infeasible when the study areas grow beyond small to medium sized regions. In order to automate the laborious manual digitzation process, some studies explored computer vision methods like trend analyses combined with random forests [8], or graph-based analysis [25].

With deep learning becoming an indispensable tool in remote sensing, it was also used for the detection of RTS features. Huang et al. [11] adapted the DeepLab architecture for semantic segmentation [26] to the task of mapping permafrost features like RTSs using imagery from unmanned aerial vehicles (UAVs) over the northeastern Tibetan Plateau. Similarly, Nitze et al. [12] trained several CNN architectures on PlanetScope satellite imagery for six study sites in northwest Canada and the Russian Arctic. Yang et al [15] combine Maxar imagery with other information like NDVI derived from Sentinel-2 and elevation information to train a CNN model to detect RTS. Huang et al. [21] opted to detect RTS directly in elevation maps instead, training an object detector on the ArcticDEM data product.

Existing studies usually focus on a single region of interest, like the Canadian Arctic [13], the Tibetan Plateau [11], [27], or a few selected regions [8], [12], [15]. More recently, efforts towards a pan-Arctic RTS data product have gained traction [21].

Other permafrost features can also be mapped using remote sensing techniques, including thermokarst lakes [8], [28], wildfires [8], [29], and ice wedges [25], [30], [31]. These research areas face similar challenges as RTS mapping, so that approaches for these tasks can also inspire new approaches for RTS mapping.

B. Self-Supervised Representation Learning

Learning features from unlabeled images has been a highly active area of research in recent years. As acquiring images is relatively simple compared to labelling them, self-supervised methods seek to train models without any labels. Still, the features derived by such models often compare competitively to fully-supervised models in evaluations [32]–[35]

Most approaches train an image encoder to embed images to feature vectors in such a way that the embedding is invariant under certain data augmentations, meaning that perturbed versions of the same image should be represented by the same point in the embedding space [32]–[35]. A trivial solution to this goal is reached when the encoder predicts the same constant feature vector for all inputs. Therefore, the main ideas that differentiate these models lie in the way that they adress this representation collapse. SimCLR [33] employs the contrastive loss function to not only match embeddings of the same image closely in the representation space, but also push apart embeddings from different images. Building on this idea, Momentum Contrast [35] introduces a momentum encoder that updates its weights as an exponential moving average of the trained models weights. Further, a queue of embeddings is used in order to leverage a larger number of negative samples. Boostrap Your Own Latent [34] uses the momentum encoder to eliminate the need for negative samples. By carefully tuning the momentum and using a projection head, this method avoids representation collapse without using a contrastive loss.

Finally, self-distillation without labels (DINO) [32] uses a different approach to eliminate negative samples. Here, the model is tasked with defining its own classification scheme for images. Two versions of the model, called student and teacher, are trained following the self-distillation process.

For a given input image, two augmentations are generated. Out of these two augmentations, the first one is run through the teacher model. The features derived by the teacher model are then centered and re-scaled. Finally, the teacher's classification is derived by applying a softmax activation to the re-scaled outputs. Meanwhile, the second version of the image is run through the student model. Finally, the student is then trained to match the teacher's classifications with its own outputs [19]. Fig. 2 outlines the DINO training process. In the following, we will be referring to the classes automatically derived by the models as "pseudo-classes".

Naturally, one a crucial step in this setup is the assignment of parameters to the teacher model. As there are no ground-truth labels in this setup, the teacher weights are taken to be an exponential moving average (EMA) of the student weights, hence the term "self-distillation".

Other than these methods, our usecase does not require image-level features, but rather pixel-wise features. With PixelDINO, we adopt the concept of self-distillation with no labels on the pixel level.

C. Semi-supervised Semantic Segmentation in Remote Sensing

In remote sensing, many relevant tasks are semantic segmentation tasks. For each pixel, a class label needs to be predicted in order to partition the entire scene into separate

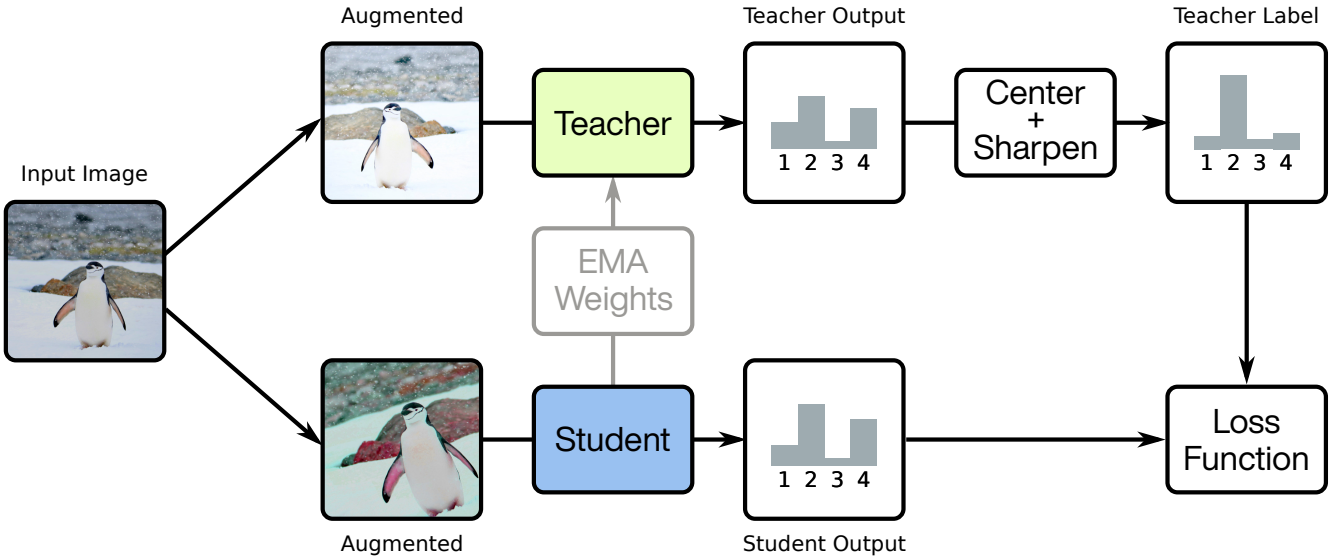


Fig. 2. Overview of the DINO framework [32] for feature learning. Two augmented versions of the input image are generated. The teacher model is then used to predict a class distribution for the first augmentation. This distribution is centered, sharpened and the softmax function is applied. The student model is then given the second augmented image and trained to predict the label given by the teacher. Finally, the teacher model's weights are updated as an exponential moving average of the student's weights.

regions of interest. Such tasks are encountered across a large number of research areas like crop type mapping [36], urban mapping [37], or monitoring animal populations [38]. Generally, it is quite hard even for experts to perfectly annotate a given scene pixel by pixel, and the process of generating these annotations is often tedious and time-consuming. There are approaches to reducing the labelling burden through working with sparse labels like point labels or scribbled labels, but these come at a price in terms of classification accuracy [39]. On the other hand, unlabeled remote sensing data is generally easily available through programmes like NASA's Landsat series or ESA's Copernicus missions. Therefore, the idea of combining small labeled datasets with large unlabeled data for semantic segmentation has been previously explored in remote sensing.

A large class of semi-supervised learning studies in remote sensing focuses on the idea of consistency regularization. The underlying assumption here is that even for unlabeled images, a model's representations or outputs should be consistent under a certain set of perturbations. For example, these perturbations can be data augmentation operations [40], feature dropout [41], additive noise in the feature space [42], [43], or interpolation between samples [44]. Under these perturbations, the model is then trained to stay consistent. This consistency can be enforced at different stages of the model calculation. Most common is the so-called pseudo-labelling technique [41], where consistency is enforced in the final output classification of the network. Various extensions of this basic idea exist [45], [46].

In FixMatch, Sohn et al. [45] enforce consistency across two sets of data augmentatations called *weak augmentations*, denoted by $\alpha(\cdot)$, and *strong augmentations*, denoted by $\mathcal{A}(\cdot)$. Upretee and Khanal [40] formulated *FixMatchSeg*, an elegant way of generalizing this framework to the semantic segmentation case. As the labels themselves are also subject to geometric transformations such as rotations, converting them between augmentations is not trivial. FixMatchSeg solves this by chaining the weak and strong data augmentations as $\mathcal{A}(\alpha(\cdot))$, so that the pseudo-label can be augmented alongside with the image.

Another possibility is to enforce consistency in the intermediate feature space within a given layer of the neural network [42]. Such approaches have been successfully applied for mapping building footprints [42], mapping landslides [47] or aerial image segmentation [48]. Our presented approach is similar to these methods. The main difference in our approach is the change from pseudo-labels to pseudo-classes. While pseudo-labels are adhering to the original classification scheme of the task, we allow the network to come up with additional classes in order to oversegment the images. This should be particularly helpful for tasks with a large class imbalance, for example when a background class with high intraclass variance dominates the scenery, which is the case in RTS detection.

The Generator-Discriminator approach from Generative Adversarial Networks (GANs) has also been explored for semi-supervised semantic segmentation. Here, the basic idea is to conceptually understand the segmentation network as either the generator or the discriminator network. In the first setup, the discriminator learns to discern true segmentation maps from model outputs on a pixel-wise level. At the same time, the segmentation network takes the role of the generator and is trained to convince the discriminator as a secondary loss objective [49]. In the other setting, a generator is used to generate synthetic data, and the discriminator is trained to differentiate these synthetic data points from the unlabeled

data, while also generating class labels [50]. Adversarial semi-supervised learning approaches have been demonstrated on tasks like hyperspectral image classification [51] or change detection [52]. Other than these works, our method only requires training a single neural network. Also, it does not exhibit the well-known training instabilities or require any of the careful hyperparameter tuning that adversarial methods are known for.

Finally, some studies separate the training process into a self-supervised pre-training phase on a large unlabeled dataset, and a supervised fine-tuning phase on the labeled dataset. As self-supervised learning has been an area of great interest in computer vision recently, this approach is getting increasingly popular. For example, such approaches have been shown to improve model performance for tasks such as hyperspectral image classification [53], land cover mapping [54], [55] or change detection [55]. Contrasting this, we present a semi-supervised training procedure where the model is trained end-to-end in a single training phase.

III. PixelDINO for Semi-Supervised Semantic Segmentation

Inspired by the ideas behind DINO [32] and FixMatch-Seg [40], we build PixelDINO, a semi-supervised semantic segmentation framework for remote sensing imagery.

A. Learning Pixel Features without Labels

While natural imagery often has a clear object of focus, a remotely sensed satellite image can have dozens or hundreds of objects of interest in it. Therefore, working on the pixel level should lead to more discriminative features, which will be crucial for a successful segmentation of these objects in the end. The main idea for our PixelDINO framework is to adopt the explained above on a pixel-wise level. Instead of classifying entire images, the student and teacher models will instead give a label to each pixel in the input image.

But In the original DINO framework, the teacher labels can be directly applied to train the student. In the pixel-wise case, data augmentations like flips or rotations will change the location of objects in the image. Therefore, pixel-wise segmentation labels also need to be augmented in the same fashion. When following the original DINO setup, doing this correctly is challenging, as it requires inverting the data augmentations applied to the first image. Further, this procedure will introduce invalid pixel labels when inverting lossy augmentations like rotations by non-multiples of 90° or cropping operators. To avoid these issues, we resort to an approach introduced by FixMatchSeg [40]. Instead of using two augmentations of the same base image, we will use a chain of augmented images.

Given an unlabeled input image $U \in \mathbb{R}^{H \times W \times C}$, we first apply a weak augmentation $\alpha(U)$ and calculate the teacher output $\mathcal{T}(\alpha(U))$. Then, the teacher's label is derived through centering, re-scaling, and applying the softmax function:

$$Y_U = \operatorname{softmax}\left(\frac{\mathcal{T}(\alpha(U)) - \mu}{\tau}\right) \tag{1}$$

Here, μ is the center of past teacher outputs, which is updated using an exponential moving average, and τ is the temperature parameter. A lower temperature leads to a stronger "sharpening" of the class distribution, which is desired in order to discourage the model from predicting a uniform distribution.

The student model $\mathcal{S}$ is applied to the strongly augmented input image to obtain the student's prediction $\mathcal{S}(\mathcal{A}(\alpha(U)))$. Finally, the PixelDINO loss is calculated as the cross entropy between the softmax of the student output and the strongly augmented teacher label:

$$\mathcal{L}_{\text{PixelDINO}} = \text{CE}(\text{softmax}(\mathcal{S}(\mathcal{A}(\alpha(U))), \mathcal{A}(Y_U)), \tag{2}$$

where CE refers to the cross-entropy operator.

In this way, the student model $\mathcal{S}$ is trained to align its predictions in such a way that they are consistent with the teacher's outputs $\mathcal{T}$ under the set of strong augmentations $\mathcal{A}$. A graphical overview of this approach is given in Fig. 3.

B. Semi-Supervised Learning with PixelDINO

The goal for semi-supervised learning is to exploit the information present in a large, unlabeled dataset and combine that with the class information from a smaller, labeled dataset. For PixelDINO, embedding the information from a labeled dataset is rather straight-forward. The DINO methodology already works with pseudo-classes, and PixelDINO extends that to pseudo-classes per pixel. If information about some specific classes is already known a priori in the form of a labeled dataset, this can be embedded into the training process in order to make the pseudo-classes align with the a priori classes. In our case, we would like to do exactly that for the RTS class from the labeled dataset.

To achieve that, we combine the PixelDINO training loop with a regular supervised training loop. In the combined training loop, the student model will be trained on both a mini-batch of labeled examples, as well as one of unlabeled examples for each training step. For a labeled example given as a pair of an image $X \in \mathbb{R}^{H \times W \times C}$ and a mask $Y \in \{0, 1\}^{H \times W}$, the supervised loss term is the regular cross-entropy which is commonly used in semantic segmentation. In practice, we also apply weak and strong data augmentation to the labeled samples:

$$\mathcal{L}_{\text{supervised}}(X, Y) = \text{CE}(\mathcal{S}(\mathcal{A}(\alpha(X))), \mathcal{A}(\alpha(Y))) \tag{3}$$

The final, semi-supervised training objective is simply the weighted sum of the two loss terms, balanced by a hyper-parameter β:

$$\mathcal{L}(X, Y, U) = \mathcal{L}_{\text{supervised}}(X, Y) + \beta \mathcal{L}_{\text{PixelDINO}}(U) \tag{4}$$

In our experiments, we find $\beta = 0.1$ to be a good choice for this hyperparameter. We analyze the influence of this hyperparameter in section V-C.

The pseudo-code for this training procedure is outlined in Alg. 1. By forcing the student model to adhere to the teacher outputs and the labeled ground truth masks at the same time, it is very likely that the classification schemes will indeed align to include one class for our desired target.

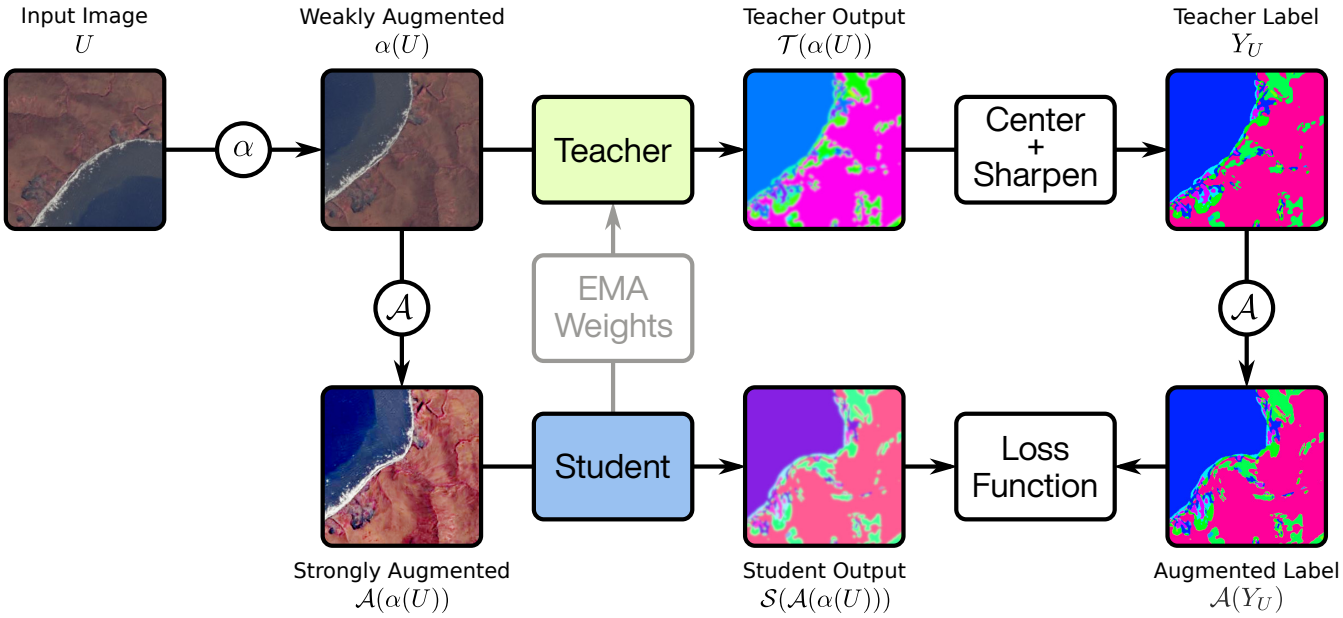


Fig. 3. Overview of the self-supervised part of the PixelDINO framework for pixel-wise feature learning. First, the image is weakly augmented and a dense feature map is derived using the teacher model. These labels are turned into class labels by centering, sharpening, and applying the softmax function. Both the weakly augmented image and the teacher label are augmented using the set of strong augmentations. The student model is then trained on this pair of image and label. Finally, the teacher model's weights are updated as an exponential moving average of the student's weights.

Algorithm 1 Semi-supervised PixelDINO (Pytorch-style)

Hyper-Parameters:
`beta`: Weight of DINO loss
`temp`: Temperature used for softmax-scaling

```
def train_step(img, mask, unlabelled):
  # Supervised Training Step
  pred = student(img)
  loss_supervised = cross_entropy(pred, mask)

  # Get pseudo-classes from teacher
  view_1 = augment_weak(unlabelled)
  mask_1 = teacher(mask_1)
  mask_1 = (mask_1 - center) / temp
  batch_center = center.mean(dim=[0,2,3])
  mask_1 = softmax(mask_1)

  # Strongly augment image and label together
  view_2, mask_2 = augment(view_1, mask_1)

  pred_2 = student(view_2)
  loss_dino = cross_entropy(pred_2, mask_2)

  loss = loss_supervised + beta*loss_dino
  loss.backward() # Back-propagate losses
  update(student) # Adam weight update
  ema_update(teacher, student) # Teacher EMA
  ema_update(center, batch_center) # Center EMA
```

C. Data Augmentations

Data augmentation is a commonly used technique to make models more robust to perturbations in the input, as well as encourage equivariance under certain geometric transformations like rotations or reflections [56]. Further, it is a crucial component for semi-supervised learning, which is why we will briefly explain the employed data augmentation techniques.

The semi-supervised learning methods introduced in this study require two different sets of data augmentation operations, in order to generate different views of the same data. Following the terminology of Sohn et al. [45], we separate the augmentations used in our study into *weak augmentations*, denoted by $\alpha(\cdot)$, and *strong augmentations*, denoted by $\mathcal{A}(\cdot)$. The conceptual difference is that weak augmentations should only add variation to the data without making the classification more difficult. Strong augmentations, on the other hand, distort the image in such a way that makes it harder for the model to perform the classification. During training, every sample is augmented randomly.

1) Weak Augmentations: In the class of weak augmentations, we only include the simple geometric transformations introduced before, namely horizontal and vertical reflections of the input imagery, as well as rotations by multiples of 90°. These augmentations are very frequently used in remote sensing as models are expected to be equivariant under reflections and rotations for most tasks.

2) Strong Augmentations: Designing a class of strong augmentations for remote sensing imagery is considerably harder than weak augmentations. The commonly used colorspace transformations which are often used for RGB imagery do not generalize well to multi-spectral imagery. Therefore, we settle for two classes of adjustments. First, we make random adjustments to the image brightness, gamma curve and contrast. In a second step, we apply rotations by arbitrary angles in the range $[-30°, 30°]$, Gaussian blurring with $\sigma = 2\text{px}$, as well as the elastic transform that locally warps parts of the image.

IV. Datasets

As the main data source for this study, we use the fourth iteration of the openly available RTS inventory from Nitze et al. [12][1]. This inventory consists of polygons that were manually labeled using PlanetScope imagery, elevation data, and Landsat timeseries as the source data. Its extent amounts to 4335 polygon annotations of RTS footprints from the years of 2018 to 2021, with a combined area of $\sim 84\,\mathrm{km}^2$. The focus of the inventory lies on multiple regions in the terrestrial Arctic, mostly in coastal areas.

While Nitze et al. [12] base their analyses on PlanetScope imagery, we opt for Sentinel-2 imagery for this study due to its open availability, which is an important factor for building a large unlabeled dataset for semi-supervised learning. Practically speaking, these two satellite platforms mainly differ in their imaging resolution and their spectral channels. While PlanetScope imagery is provided at ground sampling distances of 3–4 m and contains the visible RGB channels as well as a near-infrared channel, Sentinel-2 imagery comes at a lower spatial resolution of $10\,\mathrm{m}$ per pixel, but in turn features 13 spectral channels.

Using the image footprints from the RTS inventory, we next download 83 matching Sentinel-2 Level 1C images sourced from Google Earth Engine. As the last step, the RTS annotation polygons are rasterized to match the satellite image pixel grids. The annotation masks then contain the binary values 0 and 1 for background and RTS pixels, respectively. Similarly to Yang et al. [15], we observe good registration between the footprints and the Sentinel-2 imagery, so that no additional co-registration was performed.

Out of the annotated study regions in the original dataset, we set aside the Herschel Island and Lena sites for testing purposes. We chose the Herschel Island site for being spatially separated from the Canadian mainland. While all other study sites are in the Tundra zone, the Lena site is situated in the Boreal zone. Therefore it includes land cover features not seen in the other study sites, such as forests. This makes the Lena site a good choice for evaluating spatial generalization, leading us to choose Lena as our second test region. All of the remaining annotated regions are used as the labeled training set.

For the semi-supervised learning methods, we build a secondary unlabeled training dataset by selecting 42 Sentinel-2 tiles over permafrost areas with a focus on regions of continuous permafrost with high estimated ice content. For each one of these tiles, we then randomly select a year from the Sentinel-2 acquisition range and download the least cloudy tile taken between May and August of that year. The timespan from May to August was chosen to match the temporal distribution of the annotated data.

The obtained Sentinel-2 scenes are much larger than even modern GPU cards can handle for neural network training. Further, mini-batch training requires a uniform image size. To fulfill these requirements, all imagery is cut into patches of size 192×192 pixels as part of the training pipeline.

[1]available at https://github.com/initze/ML_training_labels

After all pre-processing steps, we arrive at a labeled training dataset with 6464 patches, an unlabeled training dataset with $266\,168$ patches, and two test datasets, Herschel and Lena, with 1052 and 4420 patches, respectively. Fig. 1 shows the spatial distribution of the labeled and unlabeled training sites.

V. Experiments & Results

A. Generalization Study

In order to quantify the improvements from the modified training procedure, we conduct experiments with different configurations. Starting with a baseline study without any training improvements, we keep the model architecture fixed and only modify the training process. For good comparability, we also use both the weak and strong data augmentations we defined in section III-C for this experiment.

Specifically, we train and evaluate models in the following configurations:

1) Baseline: Models trained only using supervised learning, without any data augmentation.

2) Baseline+Aug: Same as baseline, but trained using the weak and strong data augmentation as described in section III-C.

3) FixMatchSeg: Models trained in the semi-supervised setting using the methodology described by Upretee and Khanal [40].

4) Adversarial: Semi-supervised models trained using the adversarial approach proposed by Hung et al. [49].

5) PixelDINO: Models trained in the semi-supervised setting using our proposed methodology as outlined in Alg. 1.

As the introduced methodology focuses on adapting the training process itself rather than making changes to the model architecture, it is invariant to the specific model architecture used. Therefore, any semantic segmentation model can be used in practice. For our experiments, we use the UNet model [57] as it is a widely used network architecture for image segmentation tasks in remote sensing.

For each configuration, we train 4 models with different random seeds to also quantify the effects of the randomness in model initialization, mini-batch sampling, and data augmentation. Models were trained on a GPU server equipped with NVIDIA A6000 GPUs. The implementation was carried out in JAX [58] and Haiku [59]. The code is available online at https://github.com/khdlr/PixelDINO.

In the semi-supervised setting, the model is being trained on two datasets, the labeled data and the unlabeled data. These two datsets are vastly different in size, with the labeled dataset being much smaller than the unlabeled dataset. Therefore, the concept of "training epochs" is no longer appropriate for specifying the training duration of the model. In order to still keep comparable training schedules for the different model configurations, we instead count the number of training steps applied to each model. This should keep the comparison between the models as fair as possible, as each model has gone through the same training schedule. In all reported experiments, the models were trained for $200\,000$ steps.

TABLE II
RESULTS OF THE GENERALIZATION STUDY: MEAN AND STANDARD DEVIATION OF 4 RUNS EACH (VALUES IN %)

	Herschel					Lena				
	IoU	mIoU	F1	Precision	Recall	IoU	mIoU	F1	Precision	Recall
Baseline	19.8 ± 1.7	59.6 ± 0.9	33.0 ± 2.3	28.8 ± 3.0	39.4 ± 5.0	28.8 ± 4.0	64.3 ± 2.0	44.6 ± 5.0	52.8 ± 5.9	39.0 ± 6.0
Baseline+Aug	22.9 ± 3.0	61.3 ± 1.5	37.2 ± 3.9	44.2 ± 7.5	32.3 ± 2.0	25.8 ± 10.2	62.8 ± 5.1	40.2 ± 13.0	69.4 ± 3.2	29.4 ± 12.5
FixMatchSeg [40]	23.4 ± 0.8	61.5 ± 0.4	37.9 ± 1.1	34.1 ± 2.3	43.2 ± 4.5	32.4 ± 3.2	66.1 ± 1.6	48.8 ± 3.7	59.4 ± 2.7	41.6 ± 5.0
Adversarial [49]	26.6 ± 3.9	63.2 ± 1.9	41.9 ± 4.9	60.0 ± 9.2	32.3 ± 3.1	25.1 ± 15.1	62.4 ± 7.5	38.2 ± 20.5	87.3 ± 7.5	26.8 ± 16.7
PixelDINO	30.2 ± 2.7	65.0 ± 1.4	46.4 ± 3.2	52.7 ± 9.2	42.0 ± 3.0	39.5 ± 6.5	69.7 ± 3.3	56.4 ± 6.6	77.7 ± 6.3	44.5 ± 6.8

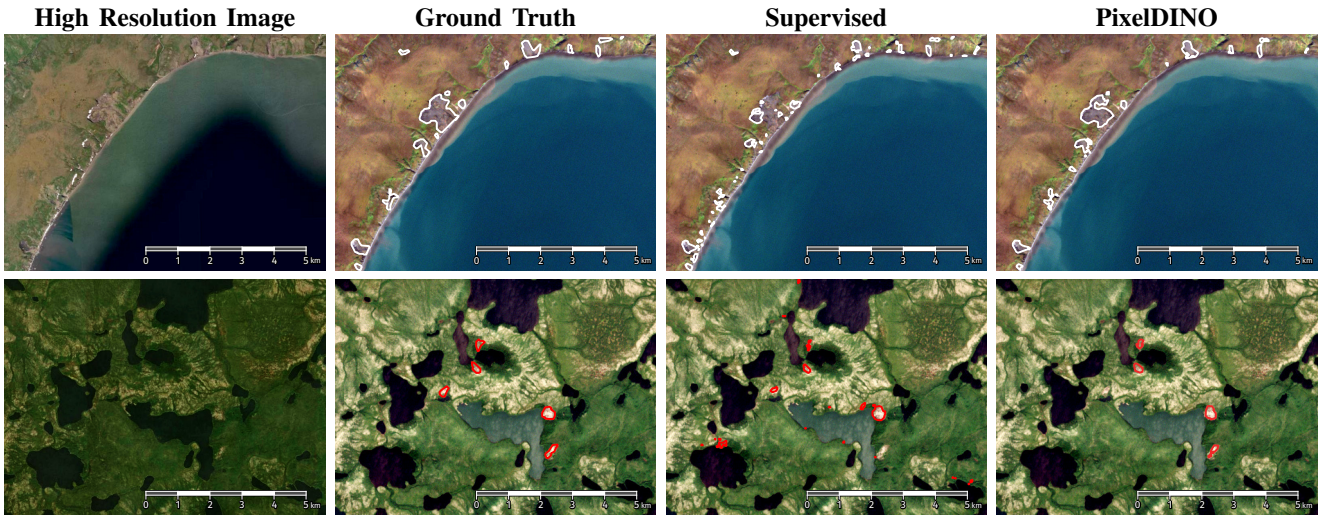


Fig. 4. High resolution imagery (1st column), ground truth (2nd column), and prediction results for parts of the Herschel Island (top) and Lena (bottom) study sites for the Baseline+Aug (3rd column) and PixelDINO (4th column) training methods. Most prominent is the large reduction in false positives due to the semi-supervised training method. The visualizations in columns 2-4 are displayed on top of Sentinel-2 data from the test datasets, high resolution imagery in column 1 courtesy of Esri, Maxar, Earthstar Geographics, and the GIS User Community.

B. Evaluation Metrics

The foreground and background classes in this dataset are highly imbalanced. Even though the study areas were chosen to feature regions of high RTS density, only around 0.7% of all pixels contain a target, while all other pixels belong to the background class. Therefore, pixel-wise accuracy is an unfit metric for this task. Instead, we evaluate the models using other metrics which are widely used for such imbalanced segmentation tasks:

1) Intersection over Union (IoU): Fraction of true positives pixels among all pixels that are true targets and/or classified positive.
2) mIoU: Mean of IoU for the RTS class and the IoU for the background class.
3) Precision: Fraction of true positive pixels among positive classifications.
4) Recall: Fraction of true positive pixels among true target pixels.
5) F1 score: The harmonic mean of Precision and Recall.

The evaluation results of the generalization study are displayed in Tab. II. Overall, the trend shows better performance of semi-supervised learning methods compared to the supervised baselines. Among the semi-supervised methods, our proposed PixelDINO approach demonstrates the strongest performance, achieving IoU scores of 30.2% for Herschel and 39.5% for Lena. The second best models score 26.6% for Herschel (Adversarial) and 32.4% for Lena (FixMatchSeg).

Although the main focus of this evaluation lies with the relative improvements from semi-supervised learning over supervised learning, we try to give an overview of how our results compare to those obtained by existing studies. Due to differences in data modalities, study regions, spatial sampling and evaluation metrics, directly comparing this study's results with existing studies is challenging. For the Herschel site, Nitze et al. [12] observe average IoU scores in the range of 20%-25% for the trained models, which is similar to the Baseline+Aug model in this study achieving an IoU of 22.9±3.0. This comparison suggests that the Sentinel-2 and Planet imagery products are comparable for RTS detection. The most comparable training setup by Yang et al. [15] is the model trained on "Extensive Sites" and evaluated on Yamal and Gydan. For this model, the study reports an mIoU of 57%, which is comparable to our baselines, which achieve mIoUs in the range of 60%-65%.

C. Influence of hyperparameter β

The PixelDINO framework introduces a tunable hyperparameter in eq. 4, namely the parameter β that determines

TABLE III
MODEL PERFORMANCE FOR DIFFERENT CHOICES OF β

	Herschel		Lena	
β	IoU	F1	IoU	F1
0.01	28.0 ± 7.3	43.4 ± 9.0	41.7 ± 2.1	58.8 ± 2.1
0.05	24.9 ± 3.6	39.7 ± 4.7	33.3 ± 2.7	49.9 ± 3.0
0.1	30.2 ± 2.7	46.4 ± 3.2	39.5 ± 6.5	56.4 ± 6.6
0.2	30.4 ± 7.7	46.2 ± 9.4	35.1 ± 15.3	50.3 ± 19.2
0.5	36.1 ± 3.8	53.0 ± 4.1	28.7 ± 15.5	42.6 ± 21.2
1.0	31.9 ± 5.3	48.2 ± 6.0	12.9 ± 3.7	22.8 ± 6.0

TABLE IV
RUNTIME OF THE EVALUATED TRAINING METHODS

Method	Training Duration	Change
Baseline	88.9 min	–
Baseline+Aug	91.3 min	+ 2.7%
FixMatchSeg	178.1 min	+ 100.3%
Adversarial	182.4 min	+ 105.2%
PixelDINO	174.9 min	+ 96.8%

the weighting of the PixelDINO loss term compared to the supervised loss term. This raises the question of how to choose the hyperparameter β. When β approaches 0, the setup becomes plain supervised learning. For very large values of β, on the other hand, the self-supervised loss term will dominate the supervised learning signal, preventing the model from learning the target classes. Intuitively, there should therefore be an optimal choice of β that balances supervised and self-supervised learning in such a way that the model performance is maximized.

We repeat our experiments for different choices of β in the range $[0.01, 1]$, the results of which are shown in Table III. Indeed, we observe that the performance generally decreases towards both edges of this interval. A choice of $\beta = 0.1$ yields good performance on both evaluation datasets. Therefore, we recommend $\beta = 0.1$ as a starting point for tuning this hyperparameter.

D. Effects on Training Duration

One common concern with increasingly complex training schemes is the increase in training time that they incur. In order to assess this, we report the average runtime of our experiments in Tab. IV. While the impact of data augmentations on the training duration is negligible, all semi-supervised training methods roughly double the duration of training. This is easily explained by the fact that the semi-supervised methods process both a batch of labeled imagery and a batch of unlabeled imagery during each iteration. However, we stress that these duration increases only occur during training and not during inference. During inference, all the presented models will run at the same speed since they share the same model architecture.

VI. DISCUSSION

The results show that for the task of RTS detection, semi-supervised learning can indeed yield a strong performance boost. In this section we will discuss our observations during the experiments, what sets apart PixelDINO from the other semi-supervised learning methods, and implications for follow-up research.

A. Isolating the Effect of Data Augmentations

As consistency across data augmentations makes up a large part of the semi-supervised training methods, the improvements in segmentation accuracy might in fact be explained by the use of data augmentations instead of the semi-supervised training itself. In order to isolate the direct effects of data augmentation on the training process, we trained the baseline supervised model with and without data augmentations.

While the data augmentations improve the model performance on the Herschel evaluation site from an IoU of 19.8% to 22.9%, they actually decrease performance for the Lena evaluation site from an IoU of 28.8% to 25.5%. This is suprising, as it is generally believed that data augmentation improves generalizability of machine learning models [56]. We attribute this to the higher land cover complexity of the Lena site, which features lakes, forest and bright bare ground and RTSs. Meanwhile, the Herschel site only featues tundra, RTSs and coastal water, matching the training data distribution more closely. Therefore, data augmentation allows the model to better detect coastal thaw slumps, while the generalization performance to inland regions suffers slightly.

At the same time, semi-supervised learning improves the performance of the baseline model much more than just applying data augmentations. From this, we conclude that the improved training performance is not explained by the data augmentations alone, but can instead be attributed to the semi-supervised learning methods.

B. Benefits of Semi-Supervised Learning

The evaluated semi-supervised methods were generally able to improve over the baselines in terms of the IoU and F1 metrics, as shown in Tab. II. Overall, semi-supervised learning has a large positive influence on the performance of the models, with the potential to increase IoU scores by around 8 basis points and F1 scores by around 12 basis points across both datasets.

The only exception here is the performance of the adversarially trained models on the Lena evaluation site. Here, this class of models actually underperforms the baselines on average. At the same time, the standard deviation is quite high, implying a large spread in model performances for this particular group. This behavior is likely tied to the most common point of criticism for adversarial training, namely that the training objective dictates a saddle-point optimization problem. These are known to be hard to solve and lead to unstable training [60]. In our experiments, this leads to unstable generalization. As the Lena test site differs much more from the training data than the Herschel site, the unstable generalization manifests itself in the Lena dataset but not in Herschel. Meanwhile, FixMatchSeg and PixelDINO do not exhibit this issue.

Generally, our proposed PixelDINO methodology achieves the strongest improvement in the segmentation metrics. This confirms that it is not only competitive with other approaches for semi-supervised semantic segmentation, but, at least for this task, is in fact the preferrable option.

C. Effects of PixelDINO Training

Our hypothesis for the strong performance of PixelDINO models lies in the fact that RTS detection is a task that has only two classes and a strong class imbalance. Therefore, the consistency regularization in approaches based on pseudo-labels like FixMatchSeg does not regularize the model sufficiently when it comes to correctly segmenting background features. This hypothesis is supported by visual inspection (see Fig. 4) and the Recall and Precision metrics in the Tab. II. While FixMatchSeg and PixelDINO have comparable Recall values, PixelDINO is far ahead in Precision, which suggests that our method is able to greatly reduce the number of false positives while maintaining a constant number of false negatives. Our findings align with Yang et al. [15], who observe that false positives are a large issue in RTS detection and address this by including negative data.

Visual inspection of the results in Fig. 4 supports our hypothesis that PixelDINO training reduces false positives. Further, while the supervised baseline sometimes fragments a single RTS target into multiple polygons, the PixelDINO predictions appear less fragmented, suggesting that our method leads to more robust predictions.

Interestingly, an inverted phenomenon can be observed for the adversarial training method. Here, the Precision values are greatly increased, beating even the models trained with PixelDINO. But this comes at the cost of poor Recall values, which means that the adversarially trained model will miss many more RTS targets than the other methods. We believe this to be related to the adversarial training method. As the discriminator is tasked with discerning true masks from predicted masks, it teaches the segmentation network mainly about the shapes of the features. While it is hard for the model to generate realistic RTS shapes, it is really easy to generate a realistic background tile by not predicting any targets. For ambiguous scenes, the adversarial model might therefore tend to predict only background, as this will always be accepted by the discriminator.

While PixelDINO appears to improve the models' robustness against false positives, we do observe slightly more false negatives in some regions, such as the Lena test set in Fig. 4. Further, as outlined in section V-D, the semi-supervised models, including PixelDINO, need roughly twice as long to train fully, as they need to ingest both unlabelled and labelled data. While the potential benefits are large, researchers therefore need to carefully consider whether the trade-offs are justified for a specific task at hand.

Overall, our PixelDINO approach greatly benefits from its ability to further subdivide the background class into regions of different semantic content, which makes the semi-supervised training feedback much more valuable, which in turn leads to more accurate predictions on the test set.

D. Avenues for Follow-Up Research

PixelDINO is easy to implement and can train more accurate RTS detectors without additional labels. We expect that these properties generalize well to other use-cases in remote sensing where data is scarce, large regional variations exist, or classes are highly imbalanced. Examples for such tasks are detecting landslides [61], flood mapping [62], or deforestation mapping [63].

It is hypothesized that satellite imagery of higher resolution will be beneficial for detecting RTSs, as oftentimes the targets can be quite small [12]. While we do not make use of such imagery due to reasons of data availability, the introduced methodology is applicable to any imagery source. It is up to future research to explore the possibilities of such methods for high-resolution satellite or even aerial imagery sources.

While not the focus of this study, a fully self-supervised version of PixelDINO might be able to learn feature maps of high spatial detail. Recent developments in foundation models [64] suggest that this is the way forward for many remote sensing tasks.

VII. Conclusion

Large volumes of remote sensing data are readily available to the public through platforms like the NASA Landsat or ESA Copernicus archives. These open up many possibile use cases for monitoring applications. Many usecases for deep learning in remote sensing are, however, hindered by a lack of sufficient labeled training data. This is particularly true for semantic segmentation tasks, because these require all pixels to be labeled. Semi-supervised learning can help relieve the labelling workload on domain experts by a large amount, simply by using readily available unlabeled data.

Our proposed PixelDINO framework achieves this by encouraging the trained model to come up with its own scheme of segmentation classes, for which it is then trained to be consistent across data augmentations as well as to align its classes to the label classes from the annotated training set.

In our experiments we demonstrated that PixelDINO can train models that generalize well to previously unseen regions in the Arctic, and do so better than both supervised baselines and other semi-supervised approaches.

As described in section VI-C, handling highly imbalanced classes is a strong property of PixelDINO. While our introduced framework is flexible in terms of the number of output channels, further research is needed to understand how well PixelDINO will generalize to semantic segmentation problems with many classes.

We expect the methods developed in this study to be transferrable to many different usecases in remote sensing even outside of permafrost monitoring. Therefore we hope to inspire follow-up research in improving the automated mapping of ground features using semi-supervised semantic segmentation methods.

Data and Code Availability

The ground truth data used in this study was published in [12] and is available at https://github.com/initze/ML_

training_labels/. The project page containing code and other materials for this study can be found at https://khdlr.github.io/PixelDINO/.

REFERENCES

[1] E. Buch, M. S. Madsen, J. She, M. Stendel, O. K. Leth, A. M. Fjæraa, and M. Rattenborg, "Arctic in situ data avilability," European Environment Agency, Kobenhavn, Denmark, Tech. Rep. 2.1, Sep. 2019.

[2] C. Gabarró, N. Hughes, J. Wilkinson, L. Bertino, A. Bracher, T. Diehl, W. Dierking, V. Gonzalez-Gambau, T. Lavergne, T. Madurell, E. Malnes, and P. M. Wagner, "Improving satellite-based monitoring of the polar regions: Identification of research and capacity gaps," in *Frontiers in Remote Sensing*, vol. 4, Feb. 2023, p. 952091.

[3] A. Bartsch, T. Strozzi, and I. Nitze, "Permafrost Monitoring from Space," *Surveys in Geophysics*, Mar. 2023.

[4] N. Nesterova, M. Leibman, A. Kizyakov, H. Lantuit, I. Tarasevich, I. Nitze, A. Veremeeva, and G. Grosse, "Review article: Retrogressive thaw slump theory and terminology," *EGUsphere*, pp. 1–36, Jan. 2024.

[5] C. R. Burn, "The thermal regime of a retrogressive thaw slump near Mayo, Yukon Territory," *Canadian Journal of Earth Sciences*, vol. 37, no. 7, pp. 967–981, Jul. 2000.

[6] H. Lantuit and W. H. Pollard, "Temporal stereophotogrammetric analysis of retrogressive thaw slumps on Herschel Island, Yukon Territory," *Natural Hazards and Earth System Sciences*, vol. 5, no. 3, pp. 413–423, May 2005.

[7] A. I. Kizyakov, S. Wetterich, F. Günther, T. Opel, L. L. Jongejans, J. Courtin, H. Meyer, A. G. Shepelev, I. I. Syromyatnikov, A. N. Fedorov, M. V. Zimin, and G. Grosse, "Landforms and degradation pattern of the Batagay thaw slump, Northeastern Siberia," *Geomorphology*, vol. 420, p. 108501, Jan. 2023.

[8] I. Nitze, G. Grosse, B. M. Jones, V. E. Romanovsky, and J. Boike, "Remote sensing quantifies widespread abundance of permafrost region disturbances across the Arctic and Subarctic," *Nature Communications*, vol. 9, no. 1, pp. 1–11, Dec. 2018.

[9] J. Hjort, D. Streletskiy, G. Doré, Q. Wu, K. Bjella, and M. Luoto, "Impacts of permafrost degradation on infrastructure," *Nature Reviews Earth & Environment*, vol. 3, no. 1, pp. 24–38, Jan. 2022.

[10] S. V. Kokelj, R. E. Jenkins, D. Milburn, C. R. Burn, and N. Snow, "The influence of thermokarst disturbance on the water quality of small upland lakes, Mackenzie Delta region, Northwest Territories, Canada," *Permafrost and Periglacial Processes*, vol. 16, no. 4, pp. 343–353, 2005.

[11] L. Huang, L. Liu, L. Jiang, and T. Zhang, "Automatic Mapping of Thermokarst Landforms from Remote Sensing Images Using Deep Learning: A Case Study in the Northeastern Tibetan Plateau," *Remote Sensing*, vol. 10, no. 12, p. 2067, Dec. 2018.

[12] I. Nitze, K. Heidler, S. Barth, and G. Grosse, "Developing and Testing a Deep Learning Approach for Mapping Retrogressive Thaw Slumps," *Remote Sensing*, vol. 13, no. 21, p. 4294, Oct. 2021.

[13] L. Huang, T. C. Lantz, R. H. Fraser, K. F. Tiampo, M. J. Willis, and K. Schaefer, "Accuracy, Efficiency, and Transferability of a Deep Learning Model for Mapping Retrogressive Thaw Slumps across the Canadian Arctic," *Remote Sensing*, vol. 14, no. 12, p. 2747, Jan. 2022.

[14] C. Witharana, M. R. Udawalpola, A. K. Liljedahl, M. K. W. Jones, B. M. Jones, A. Hasan, D. Joshi, and E. Manos, "Automated detection of retrogressive thaw slumps in the High Arctic using high-resolution satellite imagery," *Remote Sensing*, vol. 14, no. 17, p. 4132, Jan. 2022.

[15] Y. Yang, B. M. Rogers, G. Fiske, J. Watts, S. Potter, T. Windholz, A. Mullen, I. Nitze, and S. M. Natali, "Mapping retrogressive thaw slumps using deep neural networks," *Remote Sensing of Environment*, vol. 288, p. 113495, Apr. 2023.

[16] Brown, J. A. H. J., O. Ferrians, and E. Melnikov., "Circum-arctic map of permafrost and ground-ice conditions, version 2," 2002.

[17] J. E. van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Machine Learning*, vol. 109, no. 2, pp. 373–440, Feb. 2020.

[18] L. Ericsson, H. Gouk, C. C. Loy, and T. M. Hospedales, "Self-Supervised Representation Learning: Introduction, advances, and challenges," *IEEE Signal Processing Magazine*, vol. 39, no. 3, pp. 42–62, May 2022.

[19] M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 9630–9640.

[20] A. Brooker, R. H. Fraser, I. Olthof, S. V. Kokelj, and D. Lacelle, "Mapping the Activity and Evolution of Retrogressive Thaw Slumps by Tasselled Cap Trend Analysis of a Landsat Satellite Image Stack," *Permafrost and Periglacial Processes*, vol. 25, no. 4, pp. 243–256, 2014.

[21] L. Huang, M. J. Willis, G. Li, T. C. Lantz, K. Schaefer, E. Wig, G. Cao, and K. F. Tiampo, "Identifying active retrogressive thaw slumps from ArcticDEM," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 205, pp. 301–316, Nov. 2023.

[22] P. Bernhard, S. Zwieback, S. Leinss, and I. Hajnsek, "Mapping Retrogressive Thaw Slumps Using Single-Pass TanDEM-X Observations," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 3263–3280, 2020.

[23] R. A. Segal, T. C. Lantz, and S. V. Kokelj, "Acceleration of thaw slump activity in glaciated landscapes of the Western Canadian Arctic," *Environmental Research Letters*, vol. 11, no. 3, p. 034025, Mar. 2016.

[24] M. Leibman, N. Nesterova, and M. Altukhov, "Distribution and Morphometry of Thermocirques in the North of West Siberia, Russia," *Geosciences*, vol. 13, no. 6, p. 167, Jun. 2023.

[25] T. Rettelbach, M. Langer, I. Nitze, B. Jones, V. Helm, J.-C. Freytag, and G. Grosse, "A Quantitative Graph-Based Approach to Monitoring Ice-Wedge Trough Dynamics in Polygonal Permafrost Landscapes," *Remote Sensing*, vol. 13, no. 16, p. 3098, Jan. 2021.

[26] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[27] L. Huang, J. Luo, Z. Lin, F. Niu, and L. Liu, "Using deep learning to map retrogressive thaw slumps in the Beiluhe region (Tibetan Plateau) from CubeSat images," *Remote Sensing of Environment*, vol. 237, p. 111534, Feb. 2020.

[28] L. Hughes-Allen, F. Bouchard, A. Séjourné, G. Fougeron, and E. Léger, "Automated Identification of Thermokarst Lakes Using Machine Learning in the Ice-Rich Permafrost Landscape of Central Yakutia (Eastern Siberia)," *Remote Sensing*, vol. 15, no. 5, p. 1226, Jan. 2023.

[29] C. M. Gibson, L. E. Chasmer, D. K. Thompson, W. L. Quinton, M. D. Flannigan, and D. Olefeldt, "Wildfire as a major driver of recent permafrost thaw in boreal peatlands," *Nature Communications*, vol. 9, no. 1, p. 3041, Aug. 2018.

[30] C. J. Abolt, M. H. Young, A. L. Atchley, and C. J. Wilson, "Brief communication: Rapid machine-learning-based extraction and measurement of ice wedge polygons in high-resolution digital elevation models," *The Cryosphere*, vol. 13, no. 1, pp. 237–245, Jan. 2019.

[31] C. Witharana, M. A. E. Bhuiyan, A. K. Liljedahl, M. Kanevskiy, T. Jorgenson, B. M. Jones, R. Daanen, H. E. Epstein, C. G. Griffin, K. Kent, and M. K. Ward Jones, "An Object-Based Approach for Mapping Tundra Ice-Wedge Polygon Troughs from Very High Spatial Resolution Optical Satellite Imagery," *Remote Sensing*, vol. 13, no. 4, p. 558, Jan. 2021.

[32] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised Learning of Visual Features by Contrasting Cluster Assignments," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 9912–9924.

[33] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th international conference on machine learning, ICML 2020, 13-18 july 2020, virtual event*, ser. Proceedings of machine learning research, vol. 119. PMLR, 2020, pp. 1597–1607.

[34] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21 271–21 284, 2020.

[35] K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick, "Momentum contrast for unsupervised visual representation learning," in *2020 IEEE/CVF conference on computer vision and pattern recognition, CVPR 2020, seattle, WA, USA, june 13-19, 2020*. IEEE, 2020, pp. 9726–9735.

[36] L. Kondmann, A. Toker, M. Rußwurm, A. Camero, D. Peressuti, G. Milcinski, P.-P. Mathieu, N. Longépé, T. Davis, G. Marchisio, L. Leal-Taixé, and X. X. Zhu, "DENETHOR: The dynamicearthNET dataset for harmonized, inter-operable, analysis-ready, daily crop monitoring from space," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. [Online]. Available: https://openreview.net/forum?id=uUa4jNMLjrL

[37] M. Volpi and D. Tuia, "Dense Semantic Labeling of Subdecimeter Resolution Images With Convolutional Neural Networks," *IEEE Transactions*

on Geoscience and Remote Sensing, vol. 55, no. 2, pp. 881–893, Feb. 2017.

[38] E. Bowler, P. T. Fretwell, G. French, and M. Mackiewicz, "Using Deep Learning to Count Albatrosses from Space: Assessing Results in Light of Ground Truth Uncertainty," *Remote Sensing*, vol. 12, no. 12, p. 2026, Jan. 2020.

[39] Y. Hua, D. Marcos, L. Mou, X. X. Zhu, and D. Tuia, "Semantic Segmentation of Remote Sensing Images With Sparse Annotations," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[40] P. Upretee and B. Khanal, "FixMatchSeg: Fixing FixMatch for Semi-Supervised Semantic Segmentation," Aug. 2022.

[41] D.-H. Lee *et al.*, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on Challenges in Representation Learning, ICML*, vol. 3. Atlanta, 2013, p. 896.

[42] Q. Li, Y. Shi, and X. X. Zhu, "Semi-Supervised Building Footprint Generation With Feature and Output Consistency Training," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–17, 2022.

[43] L. Yang, L. Qi, L. Feng, W. Zhang, and Y. Shi, "Revisiting Weak-to-Strong Consistency in Semi-Supervised Semantic Segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7236–7246.

[44] V. Verma, K. Kawaguchi, A. Lamb, J. Kannala, A. Solin, Y. Bengio, and D. Lopez-Paz, "Interpolation consistency training for semi-supervised learning," *Neural Networks*, vol. 145, pp. 90–106, 2022.

[45] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "FixMatch: Simplifying semi-supervised learning with consistency and confidence," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, Virtual*, 2020.

[46] B. Zhang, Y. Zhang, Y. Li, Y. Wan, H. Guo, Z. Zheng, and K. Yang, "Semi-Supervised Deep learning via Transformation Consistency Regularization for Remote Sensing Image Semantic Segmentation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, pp. 1–15, 2022.

[47] F. Zhang, Y. Shi, Q. Xu, Z. Xiong, W. Yao, and XX. Zhu, "On the generalization of the semantic segmentation model for landslide detection," *Shandong Technol. Bus. Univ., Yantai, China, Tech. Rep*, 2022.

[48] J. Wang, C. H. Q. Ding, S. Chen, C. He, and B. Luo, "Semi-Supervised Remote Sensing Image Semantic Segmentation via Consistency Regularization and Average Update of Pseudo-Label," *Remote Sensing*, vol. 12, no. 21, p. 3603, Jan. 2020.

[49] W.-C. Hung, Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin, and M.-H. Yang, "Adversarial learning for semi-supervised semantic segmentation," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

[50] N. Souly, C. Spampinato, and M. Shah, "Semi supervised semantic segmentation using generative adversarial network," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5688–5696.

[51] Z. He, H. Liu, Y. Wang, and J. Hu, "Generative Adversarial Networks-Based Semi-Supervised Learning for Hyperspectral Image Classification," *Remote Sensing*, vol. 9, no. 10, p. 1042, Oct. 2017.

[52] J. Liu, K. Chen, G. Xu, H. Li, M. Yan, W. Diao, and X. Sun, "Semi-Supervised Change Detection Based on Graphs with Generative Adversarial Networks," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, Jul. 2019, pp. 74–77.

[53] N. A. A. Braham, L. Mou, J. Chanussot, J. Mairal, and X. X. Zhu, "Self Supervised Learning for Few Shot Hyperspectral Image Classification," in *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, Jul. 2022, pp. 267–270.

[54] K. Heidler, L. Mou, D. Hu, P. Jin, G. Li, C. Gan, J.-R. Wen, and X. X. Zhu, "Self-supervised audiovisual representation learning for remote sensing data," *International Journal of Applied Earth Observation and Geoinformation*, vol. 116, p. 103130, Feb. 2023.

[55] O. Mañas, A. Lacoste, X. Giró-i-Nieto, D. Vazquez, and P. Rodríguez, "Seasonal Contrast: Unsupervised Pre-Training From Uncurated Remote Sensing Data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9414–9423.

[56] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Seattle, WA, USA: IEEE, Jun. 2020, pp. 3008–3017.

[57] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, Oct. 2015, pp. 234–241.

[58] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang, "JAX: Composable transformations of Python+NumPy programs," 2018.

[59] T. Hennigan, T. Cai, T. Norman, and I. Babuschkin, "Haiku: Sonnet for JAX," 2020.

[60] D. Saxena and J. Cao, "Generative Adversarial Networks (GANs): Challenges, Solutions, and Future Directions," *ACM Computing Surveys*, vol. 54, no. 3, pp. 63:1–63:42, May 2021.

[61] P. Li, Y. Wang, G. Xu, and L. Wang, "LandslideCL: towards robust landslide analysis guided by contrastive learning," *Landslides*, vol. 20, no. 2, pp. 461–474, Feb. 2023.

[62] A. Shastry, E. Carter, B. Coltin, R. Sleeter, S. McMichael, and J. Eggleston, "Mapping floods from remote sensing data and quantifying the effects of surface obstruction by clouds and vegetation," *Remote Sensing of Environment*, vol. 291, p. 113556, 2023.

[63] A. Jamali, S. K. Roy, J. Li, and P. Ghamisi, "TransU-Net++: Rethinking attention gated TransU-Net for deforestation mapping," *International Journal of Applied Earth Observation and Geoinformation*, vol. 120, p. 103332, Jun. 2023.

[64] X. X. Zhu, Z. Xiong, Y. Wang, A. J. Stewart, K. Heidler, Y. Wang, Z. Yuan, T. Dujardin, Q. Xu, and Y. Shi, "On the foundations of earth and climate foundation models," *arXiv preprint arXiv:2405.04285*, 2024.

B Related Publications

The following publications are related to the dissertation project. While they are not direct constituents of this thesis, they represent intermediate results or originate from collaborations started as part of this dissertation project.

- **K. Heidler**, L. Mou, D. Hu, P. Jin, G. Li, C. Gan, J.-R. Wen, and X. X. Zhu, "Self-supervised audiovisual representation learning for remote sensing data," *International Journal of Applied Earth Observation and Geoinformation*, vol. 116, p. 103 130, 2023. DOI: 10.1016/j.jag.2022.103130
- E. Loebel, M. Scheinert, M. Horwath, **K. Heidler**, J. Christmann, L. D. Phan, A. Humbert, and X. X. Zhu, "Extracting glacier calving fronts by deep learning: The benefit of multispectral, topographic, and textural input features," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022. DOI: 10.1109/TGRS.2022.3208454
- E. Loebel, M. Scheinert, M. Horwath, A. Humbert, J. Sohn, **K. Heidler**, C. Liebezeit, and X. X. Zhu, "Calving front monitoring at sub-seasonal resolution: A deep learning application to Greenland glaciers," *The Cryosphere Discussions*, pp. 1–21, 2023. DOI: 10.5194/tc-2023-52
- C. A. Baumhoer, A. J. Dietz, **K. Heidler**, and C. Kuenzer, "IceLines – A new data set of Antarctic ice shelf front positions," *Scientific Data*, vol. 10, no. 1, p. 138, 2023. DOI: 10.1038/s41597-023-02045-x
- **K. Heidler**, L. Mou, C. Baumhoer, A. Dietz, and X. X. Zhu, "HED-UNet: A multi-scale framework for simultaneous segmentation and edge detection," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, 2021, pp. 3037–3040. DOI: 10.1109/IGARSS47720.2021.9553585
- **K. Heidler**, L. Mou, and X. X. Zhu, "Seeing the bigger picture: Enabling large context windows in neural networks by combining multiple zoom levels," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, 2021, pp. 3033–3036. DOI: 10.1109/IGARSS47720.2021.9554434
- **K. Heidler**, L. Mou, E. Loebel, M. Scheinert, S. Lefèvre, and X. X. Zhu, "Deep active contour models for delineating glacier calving fronts," in *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, 2022, pp. 4490–4493. DOI: 10.1109/IGARSS46834.2022.9884819